Data Quality Mining: New Research Directions

Laure Berti-Équille, Tamraparni Dasu

Data Quality Mining: New Research Directions

Laure Berti-Équille

University of Rennes 1, France

berti@irisa.fr

Tamraparni Dasu

AT&T Labs-Research, NJ, USA

tamr@research.att.com

ICDM 2009, Miami,
December 7, 2009

Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

Part III. Case Study and New Directions

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits

What is Low Data Quality?

- Missing data
- Erroneous data
- Data anomalies
- Duplicates
- Inconsistent data
- Out-of-date data
- Undocumented data

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits

Example 1

Data quality problems in a relational DB

ICDM Steering Committee

Non-standard representation

Name	*Affiliation*	*City, State, Zip, Country*	*Phone*
Piatetsky-Shapiro G.,PhD	U. of Massachusetts		617-264-9914
David J. Hand	Imperial College	London, UK	
Benjamin W. Wah	Univ. of Illinois	IL 61801, USA	(217) 333-6903
Hand D.J.			
Vippin Kumar	U. of Minnesota, MI, USA		
Xindong Wu	U. of Vermont	Burlington-4000 USA	
Philip S. Yu	U. of Illinois	Chicago IL, USA	999-999-9999
Osmar R. Zaiiane	U. of Alberta	CA	111-111-1111

Duplicates

Typos

Misfielded Value

Inconsistency

Obsolete Value

Missing Value

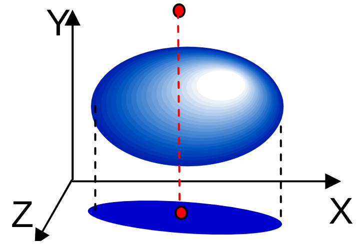
Incorrect Value

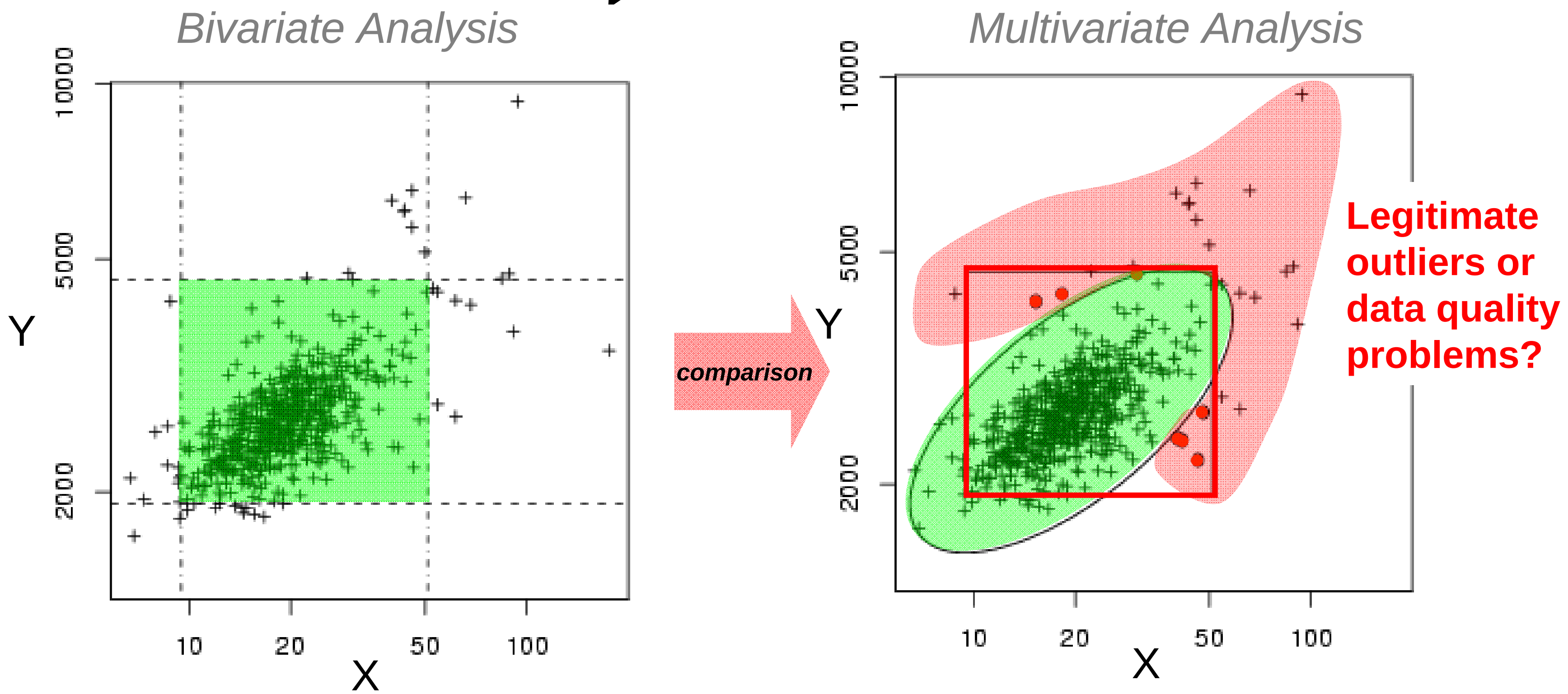
3 records are missing !

Ramamohanarao Kotagiri, U. of Melbourne, Australia
Heikki Mannila, U. of Helsinki, Finland
Shusaku Tsumoto, Shimane Univ., Japan

Incomplete Value

Example 2
Outliers





Rejection area: Data space excluding the area defined between 2% and 98% quantiles for X and Y

Rejection area based on:

Mahalanobis_dist(cov(X,Y)) > χ^2(.98,2)

Example 3

Disguised missing data

Some are obvious...

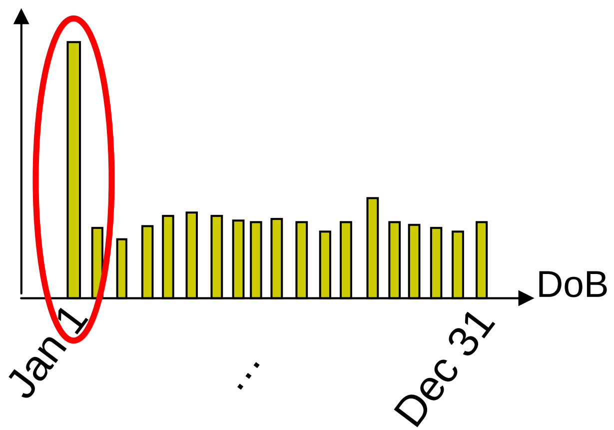
Detectable with syntactical or domain constraints

Phone number: ***999-999-9999***

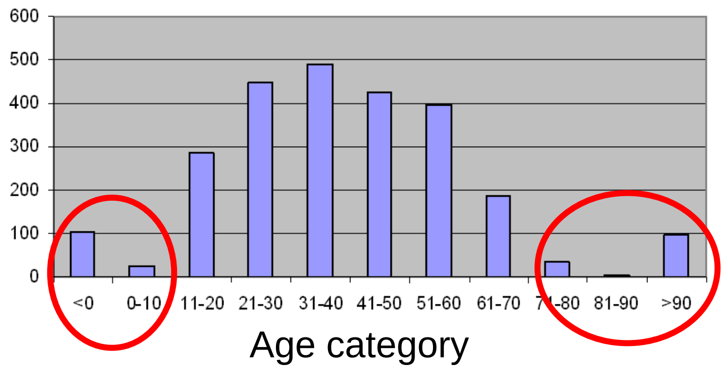
Others are not….

Could be suspected because the data distribution doesn't conform to the expected model

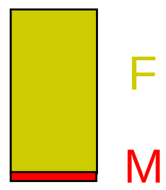
Histogram of DoBs per day of the year



Histogram of online shopping customers per age category



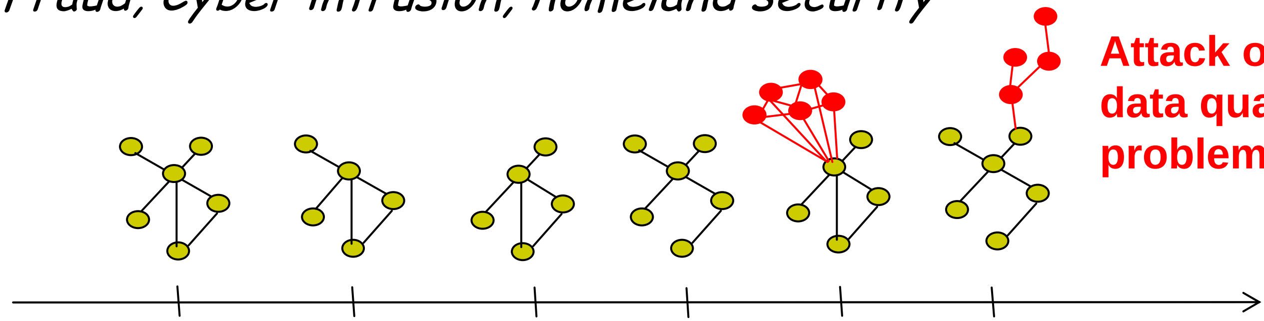
2% patients in the obstetrical emergency service are male…



Example 4

Time-Dependent Anomalies :
Unusual patterns in graph analysis

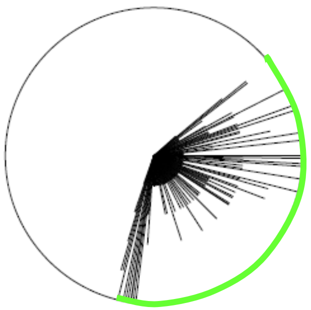
e.g., Fraud, Cyber intrusion, homeland security



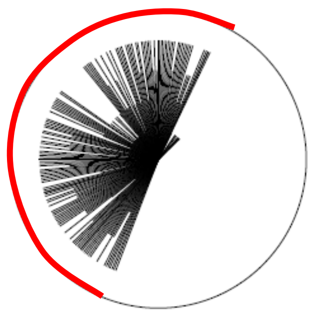
Attack or data quality problem?

e.g., IP Address Scan Patterns for a big server

Normal Scan Pattern

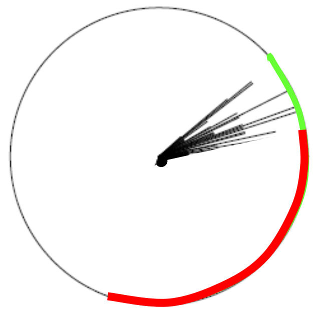


Abnormal Scan Pattern



High volume communications
with unknown IP addresses

Abnormal Scan Pattern



Data loss due to
transmission problems

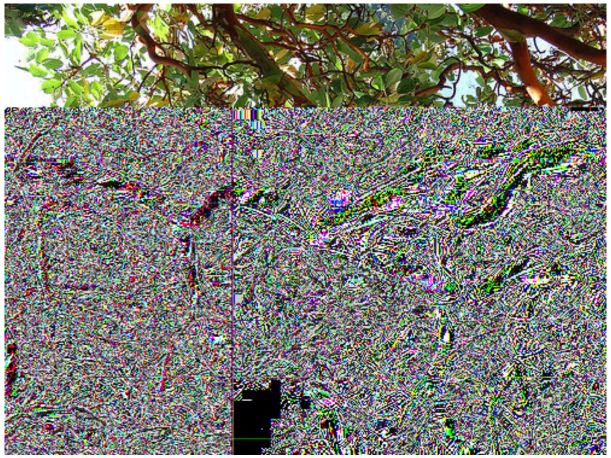
Example 5

Contradictions between Images and Text

flickr **Abuse of tags**

Arbutus tree

ADD TO FAVES BLOG THIS ALL SIZES



Tags

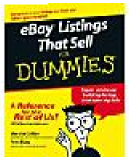
arbutus
tree
galiano
island
amsterdam
animal
animals
april
architecture
art

[…]

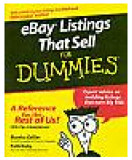
park
party
people
phone
photo
pink
portrait
red
reflection
river
roadtrip
rock
rome
sanfrancisco
school
scotland
sea
seattle
sign
sky
snow
spain
spring
street
summer
sun
sunset
taiwan
texas
thailand
tokyo
toronto
travel
trees
trip
uk
unfound
urban
usa
vacation
vancouver
washington
water
wedding
white
winter
yellow
zoo

Duplicates

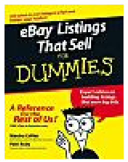
ebay **Fraud**



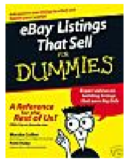
eBay Listings That Sell for Dummies - Collier, Mars New
Current Price: £11.72
View similar items...



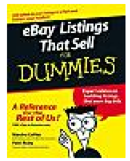
EBAY Listings That Sell Dummies E-Commerce B VALUE BOOK
Current Price: £12.89
View similar items...



EBay Listings That Sell for Dummies Book | Marsha Colli
Current Price: £13.19
View similar items...



Ebay Listings that sell for Dummies On CD- Cheap Book
Current Price: £2.99
View similar items...



eBay Listings That Sell For Dummies
Current Price: £13.19
View similar items...

Guide ID: 10000000001248134 created: 29/06/06 (updated 19/02/09) from http://reviews.ebay.co.uk/Item-Picture-Description-Theft-Prevent-eBay-Fraud_W0QQugidZ10000000001248134

Example 6

False information

Telegraph.co.uk

Home | News | Sport | Finance | Comment | Travel | Lifestyle | Culture | Fas

UK | World | Politics | Celebrities | Obituaries | Weird | Earth | Science | Health News | Educatio

HOME > NEWS > NEWS TOPICS > HOW ABOUT THAT?

Steve Jobs obituary published by Bloomberg

An obituary of very-much-alive Apple founder Steve Jobs has been accidentally published by the respected Bloomberg business news wire.

By Matthew Moore
Last Updated: 7:05PM BST 28 Aug 2008

Steve Jobs was described as the man who 'refashioned the mobile phone' in the erroneous obituary Photo: REUTERS

The story, marked "Hold for release – Do not use", was sent in error to the news service's thousands of corporate clients.

Text Size + −
Email this article
Print this article
Share this article
91 diggs digg it

How about that?
USA
News

Pictures of the day

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits

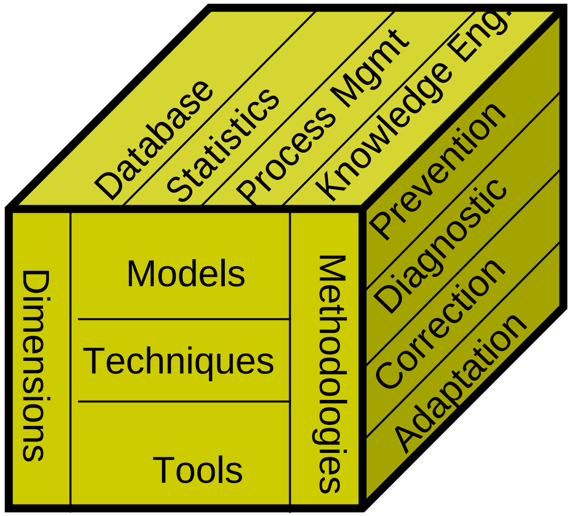
What is Data Quality?

A “subtle” combination of measurable dimensions:

- **Accuracy**
  - ICDM’09 location is in Miami Beach, France
- **Consistency**
  - Only one ICDM conference per year
- **Completeness**
  - Every past ICDM conference had a location
- **Freshness**
  - The location of the current ICDM conference is in Miami Beach
- **Uniqueness – no duplicate**
  - ICDM is a conference, not the International Confederation of Drum Manufacturers
  - ICDM’09, International Conference on Data Mining 2009 and ICDM 2009 are the same conference edition

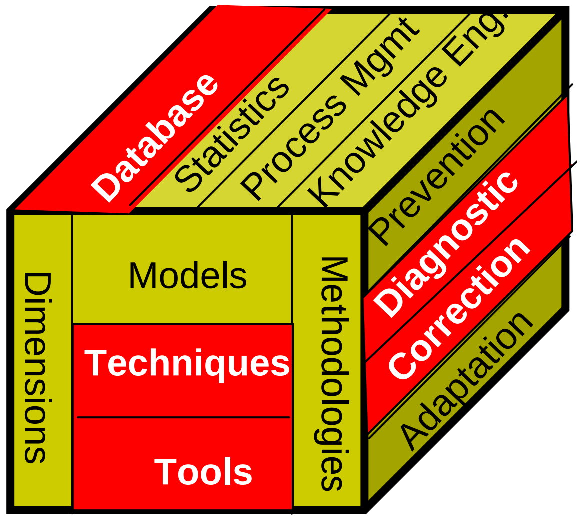
Data Quality Research:
A World of Possibilities

- 4 Disciplines
  - ✓ Statistics
  - ✓ Database
  - ✓ Knowledge Engineering
  - ✓ IT Process and Workflow Management
- 4 Types of approach
  - ✓ Prevention
  - ✓ Diagnostic
  - ✓ Correction
  - ✓ Adaptation
- 5 Levels
  - ✓ Dimensions
  - ✓ Models
  - ✓ Techniques
  - ✓ Tools
  - ✓ Methodologies



From the DB perspective

- **Data Quality Management**
  - ✓Database profiling, data auditing
  - ✓Integration of data
    - Source selection
    - Data cleaning, ETL
    - Schema and data mapping
    - Record linkage, deduplication
    - Conflict resolution, data fusion
  - ✓Constraint and integrity checking
  - ✓Data refreshment and synchronization policies
  - ✓Metadata management



From the KDD perspective

- **Data Quality Mining is beyond data preparation**

✓ Exploratory Data Analysis

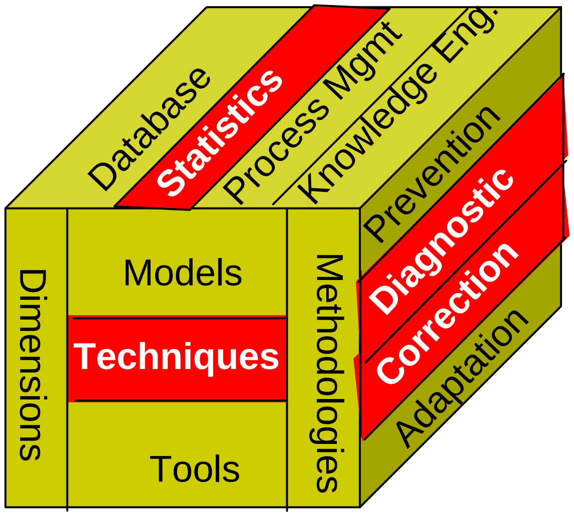
✓ Multivariate Statistics

✓ Classification
  - Rule-based
  - Model-based

✓ Clustering
  - Distance-based
  - Density-based

✓ Visualization

✓ Quantitative Data Cleaning
  - Treatment of missing values, duplicates and outliers
  - Distribution transformation



Motivation

Data quality problems are:

- Omnipresent in every application domain
- Interwoven and complex in any DB, DW or IS
- Critical to every data management, KDD and decision making project because of their massive financial impact

Limitations of current tools :

- They are *ad-hoc*, specialized, rule-based, and programmatic
- They are specific to a single-type of data quality problem
- They don’t catch interdependences between data quality problems
- Detection and cleaning tools are disconnected

Key Challenges

- Dimensionality and complexity
  - The exact notion of data quality is multidimensional and different from one application domain to another
  - Concomitant data quality problems increase the detection complexity
- Uncertainty and ambiguity
  - The boundary between quality and non-quality data is not precise
  - The boundary between a legitimate anomaly and a data quality problem is hard to define
- Dynamic
  - Data and so data quality keep evolving
- Missing Metadata

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits

Current Solutions in Practice

- Diagnostic Approaches
  - Database profiling
  - Exploratory data analysis (EDA)
- Corrective Approaches
  - Extract-Load-Transform (ETL)
  - Record linkage (RL)
  - Quantitative Cleaning

DB

Database Profiling

Include descriptive information

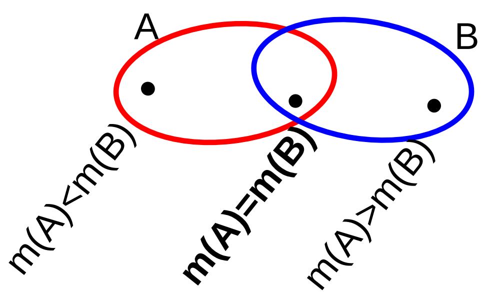
- Schema, table, domain, data sources definitions
- Business objects, rules and constraints
- Synonyms and available metadata

Systematically collect summaries of the dataset

- Number of tables, records, attributes
- Number of unique, null, distinct values for each attribute
- Skewness of data distributions
- Field Similarity (Bellman [Dasu et al., 2002])
  - By exact match
  - By substring similarity
    - Resemblance of Q-gram signatures
    - Resemblance of Q-gram min-hash distributions
- Finding Keys and FDs

Resemblance of 2 sets A and B

$\rho(A,B) = |A \cap B| / |A \cup B|$



$Pr[m(A) = m(B)] = \rho(A,B)$

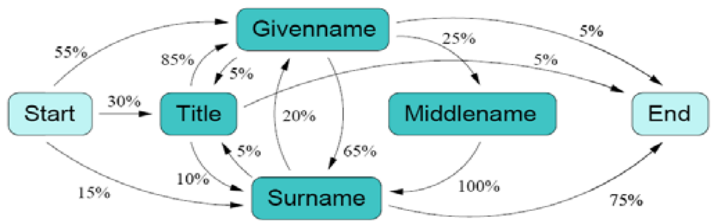
Mainly applied to relational data

Extract-Transform-Load and Cleaning

Goals

- Format conversion
- Standardization of values with loose or predictable structure
  - e.g., addresses, names, bibliographic entries
- Abbreviation enforcing
- Data consolidation based on dictionaries and constraints

Approaches



[Christen et al., 2002]

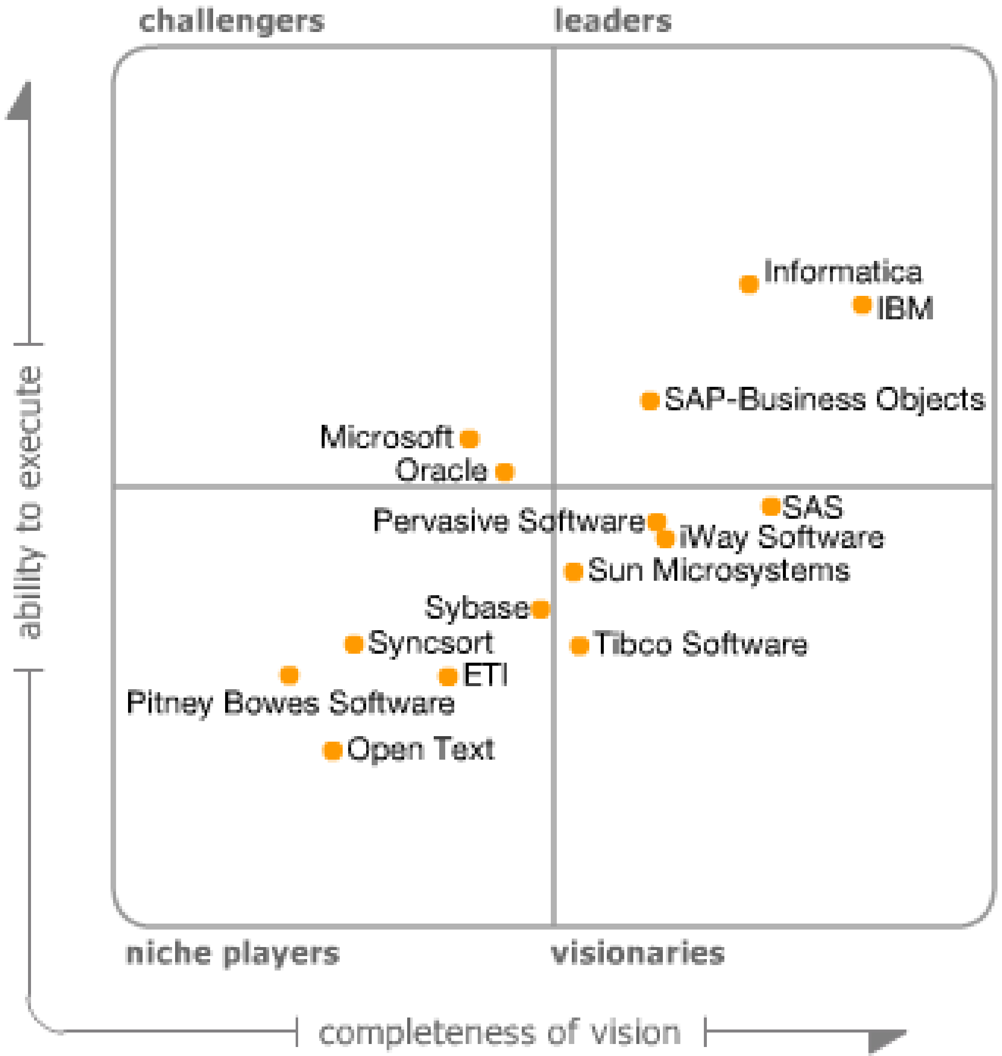
- Declarative language extensions
- Machine learning and HMM for field and record segmentation
- Constraint-based method [Fan et al., 2008]

Performance and scalability issues of most ETL tools

Academic and Open Source ETL Tools

Name	Main characteristics	Data Transformation	Data Cleaning	Duplicate Detection	Data Enrichment	Data Profiling	Data Analysis
Potter's wheel [Raman *et al.* 2001]	Detection and correction of errors with data transformations: *add, drop, merge, split, divide, select, fold, format* Interactivity, inference of the data structure	X					X
Ajax [Galhardas *et al.* 2001]	Declarative language based on logical transformation operators: *mapping, view, matching, clustering, merging* 3 algorithms for record matching	X	X	X	X		
Arktos [Vassiliadis 2000]	Graphical and declarative (SQL-like and XML-like) facilities for the definition of data transformation and cleaning tasks, optimization, measures of quality factors	X	X				
Intelliclean [Low *et al.* 2001]	Detection and correction of anomalies using a set of rules (*duplicate identification, merge, purge, stemming, soundex, stemming, abbreviation*) - Not scalable			X			
Bellman [Dasu et al., 2002]	Data quality browser collecting database profiling summaries, implementing similarity search, set resemblance, Q-gram sketches for approximate string matching			X		X	X
Febrl [Christen, 2008]	Open source in Python, initially dedicated to data standardization and probabilistic record linkage in the biomedical domain, including Q-gram, sorted NN, TF-IDF methods for record linkage and HMM-based standardization	X	X	X		X	X
Pentaho-Kettle http://kettle.pentaho.org	Open source in Java for designing graphically ETL transformations and jobs such as reading, manipulating, and writing data to and from various data sources. Linked to Weka. Easily extensible via Java Plug-ins	X	X	(X)	(X)	(X)	(X)
Talend Open Studio http://www.talend.com	Open source based on Eclipse RCP including GUI and components for business process modeling, and technical implementations of ETL and data flows mappings. Script are generated in Perl and Java code.	X	X	(X)	(X)	(X)	(X)

Commercial ETL Tools



Source: Magic Quadrant for **Data Integration Tools**, Sept. 2008, Gartner RAS Core Research Note G00160825.

Criteria

Ability to execute

- Product/Service
- Overall Viability
- Sales Execution/Pricing
- Market Responsiveness
- Track Record
- Marketing Execution
- Customer Experience
- Operations

Completeness of vision

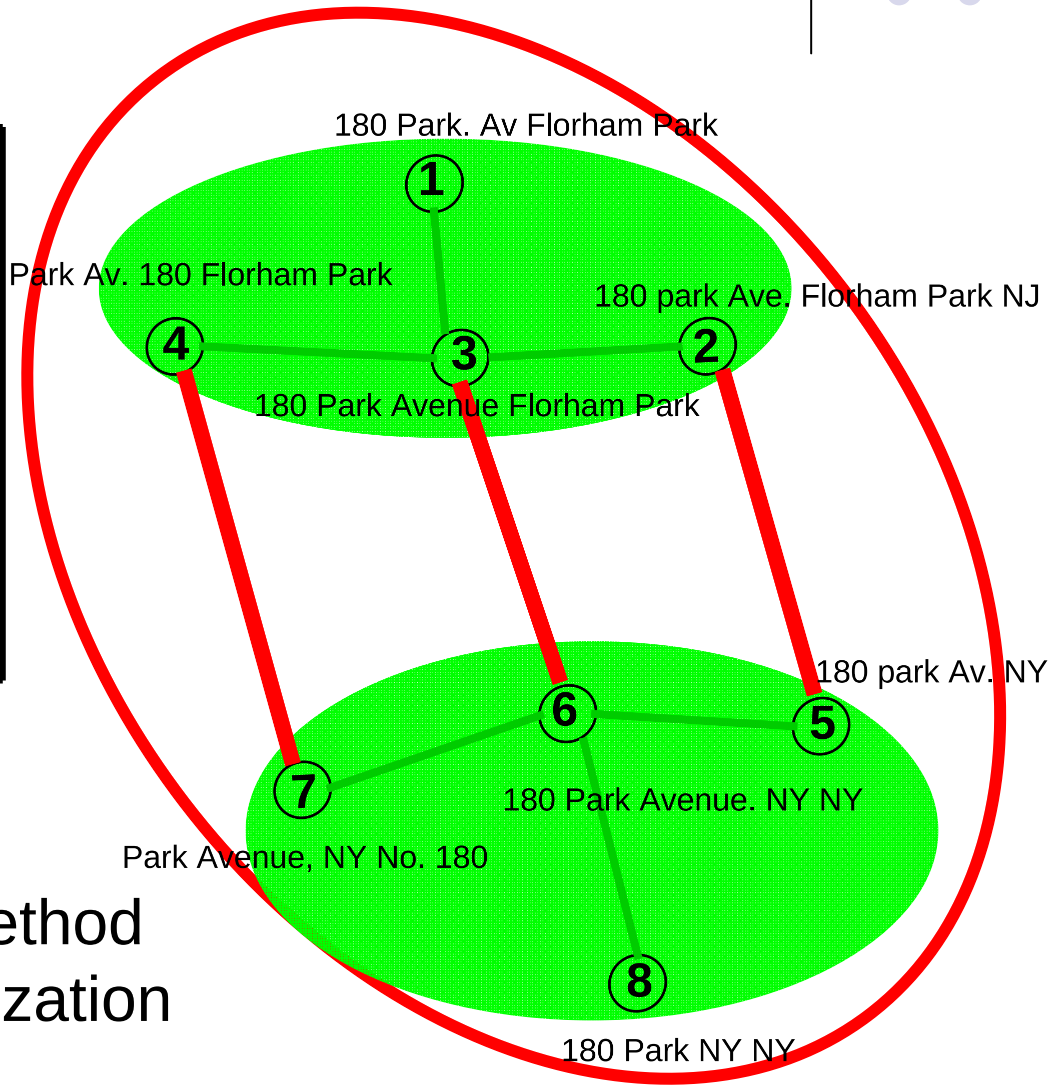
- Market Understanding
- Marketing Strategy
- Sales Strategy
- Offering (Product) Strategy
- Business Model
- Vertical/Industry Strategy
- Innovation
- Geographic Strategy

Record Linkage (RL) [Elmagarmid et al., 2007]

1. Pre-processing: transformation and standardization
2. Select a blocking method to reduce the search space partitioning the dataset into mutually exclusive blocks to compare
   - Hashing, sorted keys, sorted nearest neighbors
   - (Multiple) Windowing
   - Clustering
3. Select and compute a comparison function measuring the similarity distance between pairs of records
   - Token-based : N-grams comparison, Jaccard, TF-IDF, cosine similarity
   - Edit-based: Jaro distance, Edit distance, Levenshtein, Soundex
   - Domain-dependent: data types, ad-hoc rules, relationship-aware similarity measures
4. Select a decision model to classify pairs of records as matching, non-matching or potentially matching
5. Evaluation of the method (recall, precision, efficiency)

Chaining or Spurious Linkage

ID	*Name*	*Address*
1	AT&T	180 Park. Av Florham Park
2	ATT	180 park Ave. Florham Park NJ
3	AT&T Labs	180 Park Avenue Florham Park
4	ATT	Park Av. 180 Florham Park
5	TAT	180 park Av. NY
6	ATT	180 Park Avenue. NY NY
7	ATT	Park Avenue, NY No. 180
8	ATT	180 Park NY NY



- Expertise required for method selection and parameterization

Interactive Data Cleaning

- **D-Dupe** [Kang et al., 2008] http://www.cs.umd.edu/projects/linqs/ddupe
  Duplicate search and visualization of cluster-wise relational context for entity resolution
- **Febrl** [Christen, 2008]: https://sourceforge.net/projects/febrl/
  Rule-based and HMM-based standardization and classification-based record linkage techniques
- **SEMANDAQ** [Fan et al., 2008]: CFD-based cleaning and exploration
- **HumMer** [Bilke et al., 2005]: Data fusion with various conflict resolution strategies
- **XClean** [Weis, Manolescu, 2007]: Declarative XML cleaning

Inconsistent Data

- **Probabilistic Approximate Constraints** [Korn et al., 2003]

  Given a legal ordered domain on an attribute,

  - A **domain PAC** specifies that all attribute values x fall within ε of D with at least probability δ, as $\Pr\left(x \in \left[D \pm \varepsilon\right]\right) \geq \delta$
  - A **functional dependency PAC** X → Y specifies that, if $\left|T_i.A_\ell - T_j.A_\ell\right| \leq \Delta_\ell \;\; \forall A_\ell \in X$ then $\Pr\left(\left|T_i.B_\ell - T_j.B_\ell\right| \leq \varepsilon_\ell\right) \geq \delta \;\; \forall B_\ell \in Y$

- **Pseudo-constraints** [Ceri et al., 2007]

  Pair <*P1,P2*> where *P1* and *P2* are predicates on the same domain *D* such that if *P1* holds, then usually *P2* also and therefore there are *few* rule violations. More formally, based on the probability contingency table,

$$\frac{p_{11}}{p_{11} + p_{21}} - \rho - (1-\rho).(p_{11} + p_{12}) > 0$$

	$P1$	$\overline{P1}$	
$P2$	p_{11}	p_{12}	$p_{1.}$
$\overline{P2}$	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

- **Pattern Tableaux for Conditional Functional Dependencies**

  [Bohannon et al. 2007, Bravo et al. 2007, Golab et al. 2008, Fan et al. 2009]

  A CFD is defined to be a pair $\varphi = R(\underbrace{A \rightarrow B,}_{\textit{Embedded FD}} T_p)$, where $T_p =$

A	B
_	b_1
_	b_2

Open Issues in DQ management

Data Profiling

- Summaries refreshment
- Incremental re-computation strategies

DQ Monitoring

- Continuous checking of statistical constraints

ETL

- Extending declarative languages with constraints on DQ
- Active warehousing: online processing operators
- Optimization
- Assistance and recommendation of alternative ETL scenarios

Deduplication

- Benchmarks
- Over-matching problem
- Scalability
- Multi-objective optimization problem

Current Solutions in Practice

- Diagnostic Approaches
  - Database profiling
  - Exploratory data analysis (EDA)

KDD

- Corrective Approaches
  - Extract-Load-Transform (ETL)
  - Record linkage (RL)
  - Quantitative Cleaning

Exploratory Data Analysis (EDA)

EDA

- Use of simple statistical techniques for exploring and understanding the data
- Usually for variable and model selection and for testing distributional assumptions

EDA for Data Quality

- Detect data glitches
  - Outliers and extremes
  - Missing values
  - High frequency values and duplicates
- Data transformation for model fitting
- Treatment of glitches
  - Selecting variables and records
  - Replacing using statistical models

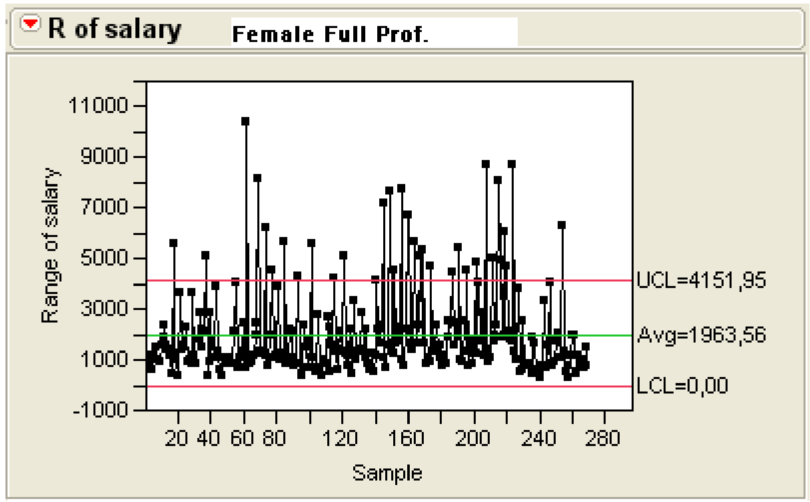
EDA – Outlier Detection

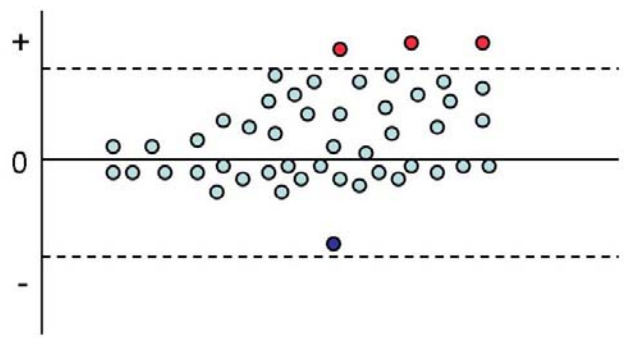
- Control chart/error bounds methods
  - e.g., expected value; confidence interval or error bounds; 3-Sigma, Hampel bounds, IQR
- Model-based outlier detection methods
  - e.g., regression model: outlyingness measured through residuals that capture deviation from the model
- Multivariate statistics for outlier detection
  - e.g., density-based and geometric or distance-based outlier detection

EDA - Control chart/error bounds

- Typical value (green) – arithmetic mean, median
- Error bounds (red) – standard deviation, IQR
- Underlying assumptions of normality and symmetry
- Simple, but potential for misleading conclusions
- Non trivial to extend to higher dimensional space

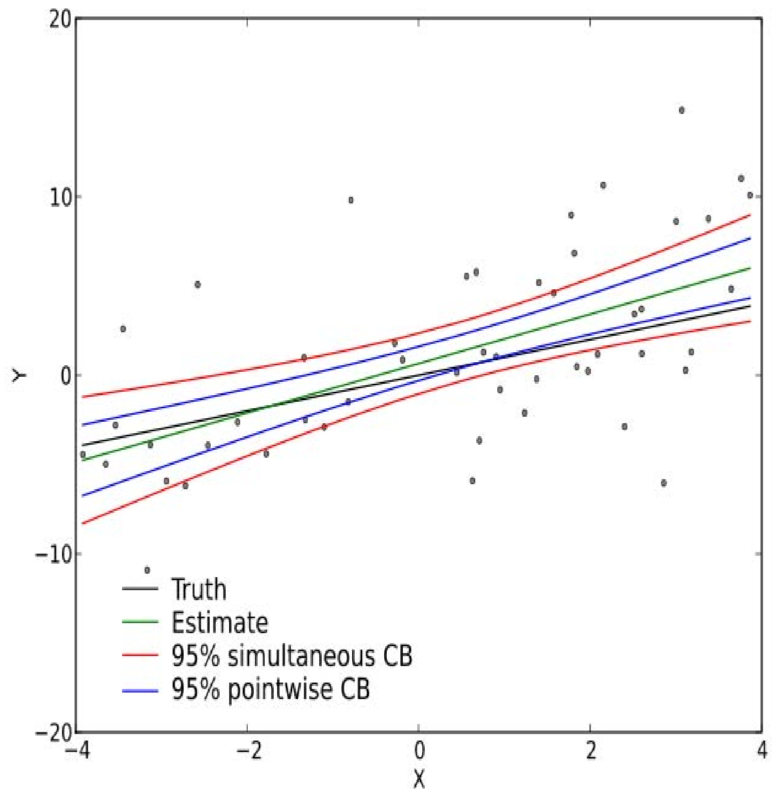
R chart





EDA - Model-based outlier detection

- Model captures relationships between variables
- Confidence bounds/bands capture variability
- Points that lie outside bounds
- The choice and correctness of the model are critical
- Expertise required for choosing the model and variables



http://en.wikipedia.org/wiki/File:Regression_confidence_band.svg

Nonparametric methods

- No obvious models?
- Projections and subspaces
  - PCA
  - Robustness
- Distance based
- Density based

Finding Multivariate Outliers

INPUT: An N x D dataset (N rows, D columns)

OUTPUT: Candidate Outliers

1. Calculate the mean μ and the D x D variance–covariance matrix Σ
2. Let C be a column vector consisting of the square of the Mahalanobis distance to the mean μ as:

$$(x-\mu)'\Sigma^{-1}(x-\mu)=(x-\mu)'\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd}\end{bmatrix}^{-1}(x-\mu)$$

c 3. Find points O in C whose value is greater than $inv\left(\sqrt{\chi_d^2(.975)}\right)$
4. Return O.

- Mean and standard deviation are extremely sensitive to outliers (Breakdown point=0%)

Robust estimators

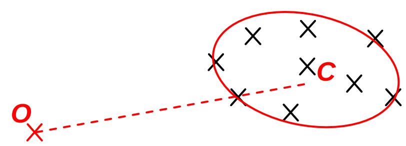
Minimum Covariance Determinant (MCD) [Rousseeuw & Driessen, 1999]

Given n data points, the MCD is the mean and covariance matrix based on the sample of size h ($h < n$) that minimizes the determinant of the covariance matrix.

Minimum Volume Ellipsoid (MVE) [Rousseeuw & Van Zomeren, 1990]

Let the column vector C with the length d ($d > 2$) be the estimate of location and let the d-by-d matrix $\mathbf{M}$ be the corresponding measure of scatter. The distance of the point $x_i = (x_{i1},\ldots,x_{id})$ from C is given by:

$$D_i = \sqrt{(x_i - C)' \mathrm{M}^{-1} (x_i - C)}$$



If $D_i > \sqrt{\chi^2_{.975,d}}$ then x_i *is declared an outlier.*

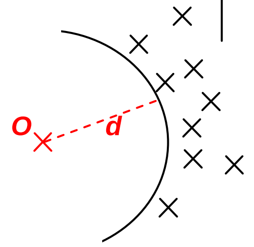
C is center of the minimum volume ellipsoid covering (at least) h points of the data set.

- Masking the structure of the group of MV outliers (clustered vs scattered)

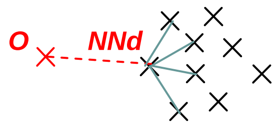
EDA - Distance-based outliers

Nearest Neighbour-based Approaches

A point *O* in a dataset is an *DB*(*p*,*d*)-outlier if at least fraction *p* of the points in the data set lies greater than distance *d* from the point *O.* [Knorr, Ng, 1998]



Outliers are the top *n* points whose distance to the *k*-th nearest neighbor is greatest. [Ramaswamy et al., 2000]



- Methods fails
  - When normal points do not have sufficient number of neighbours
  - In high dimensional spaces due to data sparseness
  - When datasets have modes with varying density
- Computationally expensive

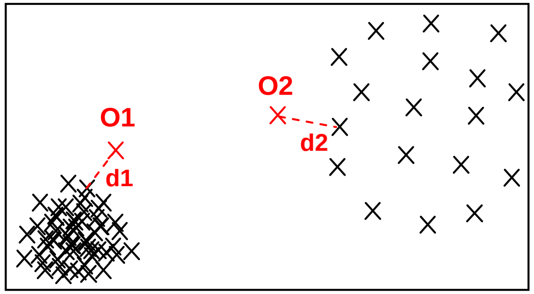
EDA - Density-based outliers

Method

Compute local densities of particular regions and declare data points in low density regions as potential anomalies

Approaches

- Local Outlier Factor (LOF) [Breunig et al., 2000]
- Connectivity Outlier Factor (COF) [Tang et al., 2002]
- Multi-Granularity Deviation Factor [Papadimitriou et al., 2003]



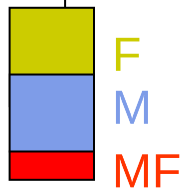
NN: O2 is outlier but O1 is not
LOF: O1 is outlier but O2 is not

- Difficult choice between methods with contradicting results
- In high dimensional spaces, factor values will tend to cluster because density is defined in terms of distance

Quantitative Data Cleaning

Methods

- **Inclusion** *(applicable for less than 15%)*
  - Anomalies are treated as a specific category



- ***Deletion***
  - List-wise deletion omits the complete record *(for less than 2%)*
  - Pair-wise deletion excludes only the anomaly value from a calculation
- ***Substitution*** *(applicable for less than 15%)*
  - Single imputation based on mean, mode or median replacement
  - Linear regression imputation
  - Multiple imputation (MI)
  - Full Information Maximum Likelihood (FIML)

Limits of EDA methods

Detection | **Cleaning** | **Explanation**

- Classical assumptions won't work

  (e.g., MCAR/MAR, normality, symmetry, uni-modality)

- DQ problems are not necessarily rare events
- DQ problems may be (partially) correlated
- Explanatory variables/processes may be external and out of reach
- **Mutual masking-effects impair the detection**

  (e.g., - missing values affects the detection of duplicates
  - duplicate records affects the detection of outliers
  - imputation methods may mask the presence of duplicates)

Limits of EDA methods

Cleaning **Explanation**

- The space of cleaning strategies is infinite
- DQ problems are domain-specific – hard to find general solutions
- Cleaning solutions may introduce new DQ problems
- Benchmarking cleaning strategies and *ad hoc* practices is hard (never been done)

What is Data Quality Mining?

"DQM can be defined as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases." [Hipp, Güntzer, Grimmer, 2001]

In addition,

Data Quality Mining (DQM) intends to be an iterative framework for creating, adapting, and applying data mining techniques for the discovery, explanation and quantitative cleaning of data glitches and their complex patterns in large and patchy datasets.

Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

Part III. Case Study

Part II. Data Quality Mining

1. Outlier Mining
2. Change Detection
3. Handling Missing and Duplicate Data

Part II. Data Quality Mining

1. Outlier Mining
2. Change Detection
3. Handling Missing and Duplicate Data

Outlier Mining

- Multivariate techniques
  - Projection pursuit
  - Distance and depth based methods
  - Probability and kernel based methods
- Stream specific methods
- Too many outliers → Distributional shift?
  - Change detection
- Great tutorial on outliers [Kriegel et al., 2009]:
  http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/tutorial_slides.pdf

Projection Based Methods

- Projection pursuit techniques are *applicable in diverse data situations* although at the expense of high computational cost.
  - No distributional assumptions, search for useful projections
- *Robust:* Filzmoser, Maronna, Werner (2008) propose a fast method based on robust PCA with differential weights to maximally separate outliers. Shyu et al. (2003) use a similar theme.
- *Time Series:* Galeano et al. (2006) extend the idea of projecting in directions of high and low kurtosis to multivariate time series.
- *Skewed Distributions:* Hubert and Van der Veeken (2007) extend the boxplot idea by defining adjusted outlyingness followed by random projections for detecting outliers in skewed data.

Outlier Mining - Robust PCA

[Shyu et al., 2003]

INPUT: An N x d dataset

OUTPUT: Candidate Outliers

1. Compute the principal components of the dataset
2. For each test point, compute its projection on these components
3. If y_i denotes the i^{th} component, then the following has a chi-square distribution

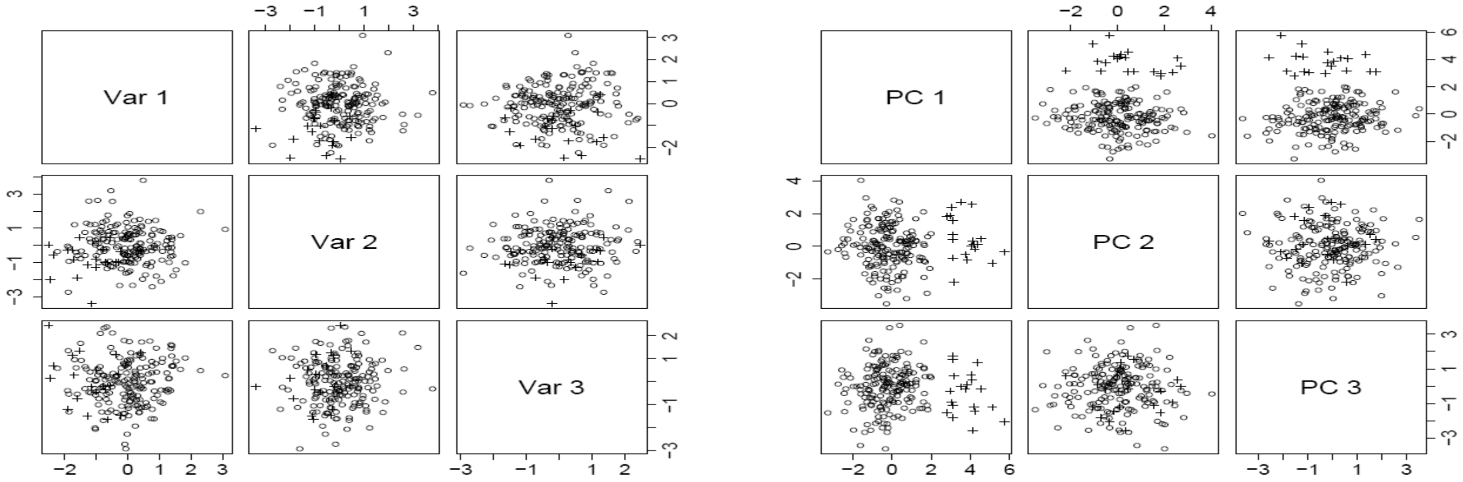
$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_q^2}{\lambda_q}, q \le p$$

3. For a given significance level α, an observation is an outlier if

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} \ge \chi_q^2(\alpha)$$

Outlier Identification in High Dimensions

[Filzmoser, Maronna and Werner, 2008]

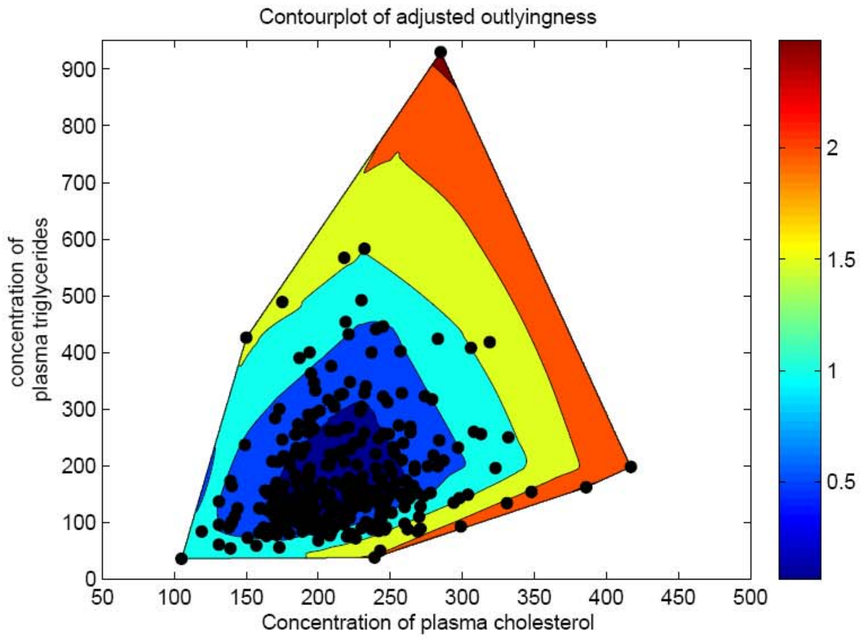


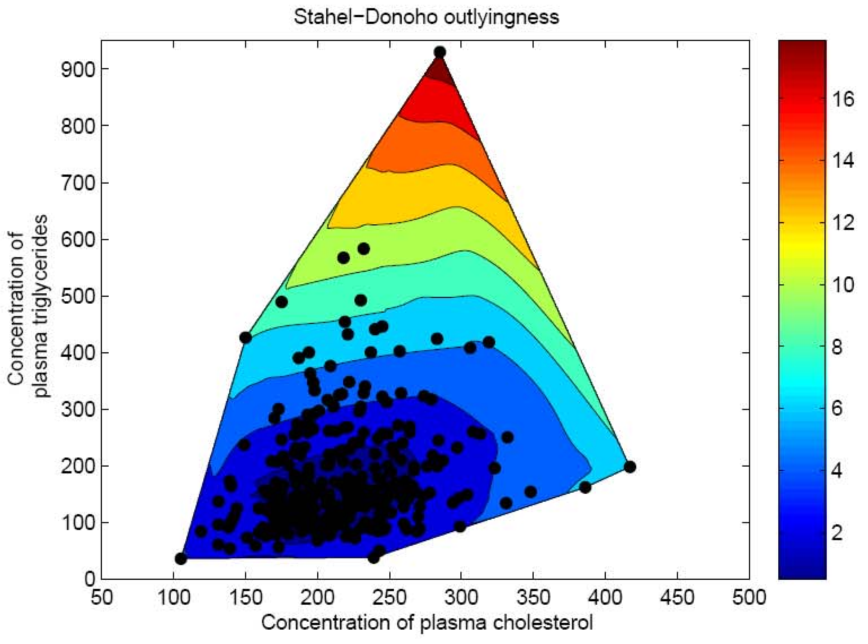
- Works in very high-D, where dimensions > samples, e.g., gene data
- Differential weights to detect location and scatter outliers; weights combined in final step
- Based on robust statistics

Outlier Detection for Skewed Data

[Hubert and Van der Veeken, 2007]

- For skewed distributions
- Key concepts
  - Adjusted outlyingness – different scaling on either side of median in boxplots.
  - MV equivalent, e.g., bagplot in 2-D
  - Random projections to identify outliers





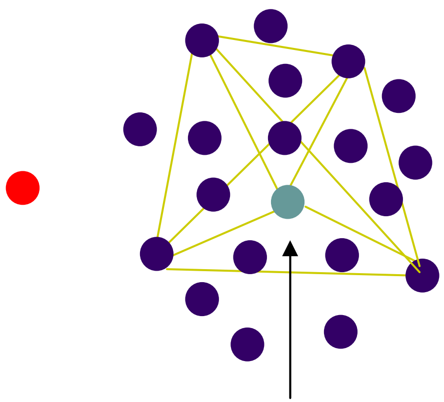
Distance and Depth Based Methods

- Distance-based methods aim to detect outliers by computing a measure of how far a particular point is from most of the data.
- *Robust methods*
  - Robust distance estimation in high-D [Maronna and Zamar, 2002] [Pena and Prieto, 2001]
- *Depth based nonparametric methods*
  - Nonparametric methods based on multivariate control charts [Liu et al, 2004]
  - Outlier detection with kernelized spatial depth function [Cheng, Dang, Peng and Bart, 2008]
- *Exotic methods*
  - Angle based detection [Kriegel, 2008]

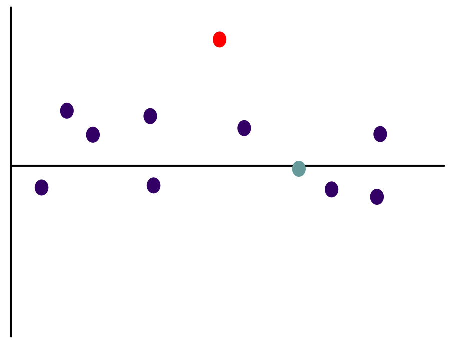
DDMA: Nonparametric Multivariate Moving Average Control Charts Based on Data Depth

[Liu, Singh and Teng, 2004]

- Extends simplicity of control charts to higher dimensions – relatively few assumptions
- Use any data depth, e.g., simplicial depth to map multidimensional data to a scalar and rank
- Apply moving average control chart techniques to data depth rank to identify outliers



$$X \in \mathrm{R}^d \rightarrow D \in \mathrm{R}$$



Deepest point, e.g., simplicial depth = contained in most triangles

Other methods

- *Popular methods*: LOF, INFLO, LOCI
  see Tutorial of [Kriegel et al., 2009]
- *Mixture distribution:* Anomaly detection over noisy data using learned probability distributions [Eskin, 2000]
- *Entropy:* Discovering cluster-based local outliers [He, 2003]
- *Projection into higher dimensional space:* Kernel methods for pattern analysis [Shawne-Taylor, Cristiani, 2005]

Probability Based Methods

- **Probability distributions** [Eskin, 2000]

  Assumption:

  High probability to have the number of normal elements in a dataset D significantly larger than the number of outliers

  Approach:

  From the distribution for the dataset D given by: $D= (1-\lambda)\, M + \lambda\, A$

  with M: Majority distribution and λ: Anomaly distribution

  - Compute likelihood of D at time t: $L_t(D)$
  - Compare $L_t(D)$ *with* $LL_t(D)$ assuming the point o_t is outlier at time t

- **Entropy-based methods** [He 2003]

  Approach:

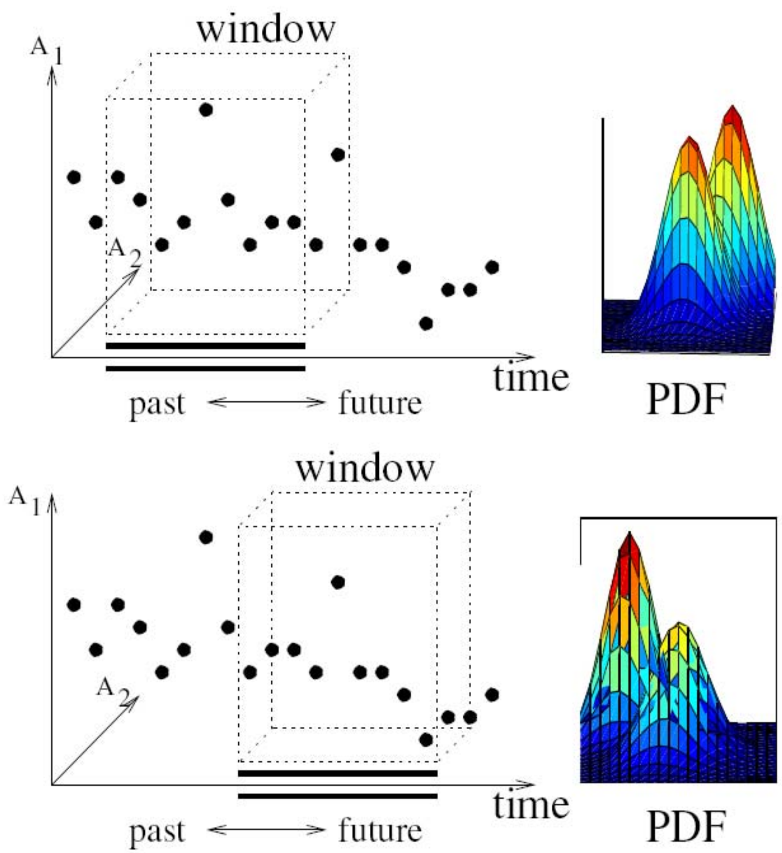
  Find a k-sized subset whose removal leads to the maximal decreasing of entropy

Stream Specific Methods

- *Distance based outliers:* Detecting distance based outliers in streams of data. [Anguilli and Fassetti, 2007]
- *Distributed streams:* Adaptive Outlier Detection in Distributed Streams [Su, Han, Yang, Zou, Jia, 2007]
- *A general density estimation scheme:* Online outlier detection in sensor streams [Subramaniam et al , 2006]
- *Projections and high dimensions*: Projected outliers in High-D data streams [Zhang, Gao, Wang, 2008]
- *Items of interest:* Finding frequent items in data streams [Cormode and Hadjieleftheriou, 2008]

Online Outlier Detection in Sensor Data Using Non-Parametric Models
[Subramaniam et al., 2006]

- Online outlier detection in hierarchical sensor networks
- Solve the more general problem of estimating the multidimensional data distribution
  - Chain sampling
  - Epanechnikov kernel

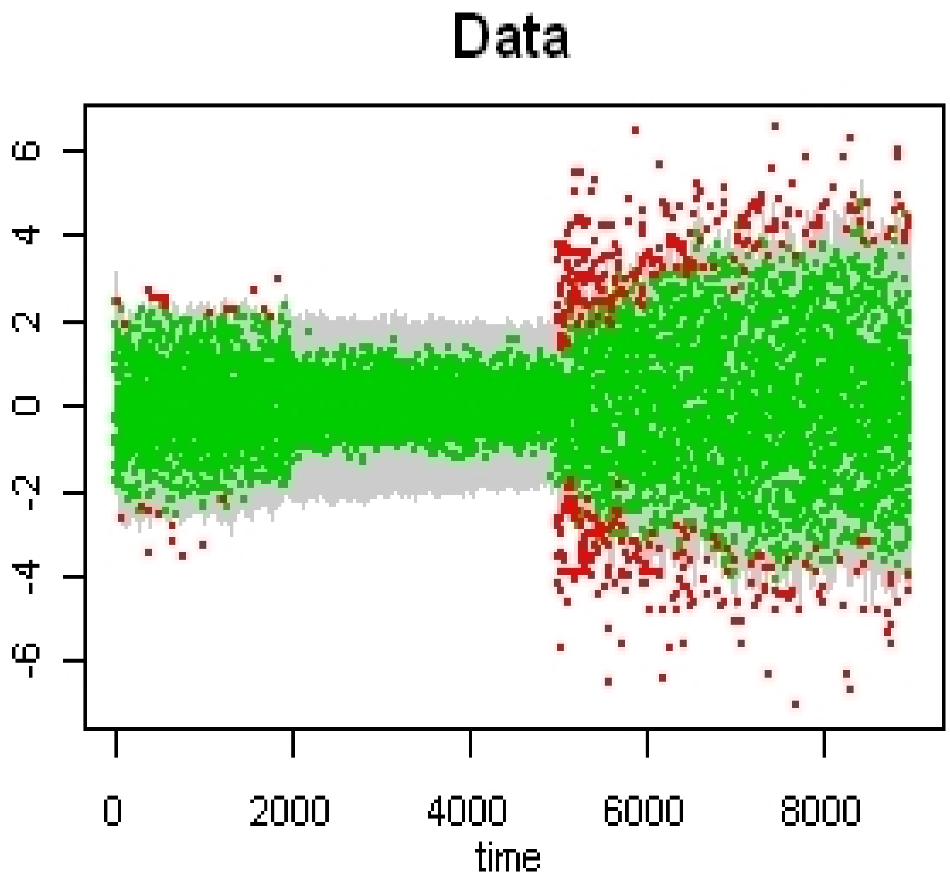


Part II. Data Quality Mining

1. Outlier Mining
2. Change Detection
3. Handling Missing and Duplicate Data

Outliers and Change Detection

- Often, an increase or decrease in outliers is the first sign of a distributional shift
- Serious implications for data quality – recalibrate anomaly detection methods
- Change detection methods are critical



Difference in Data Distributions

- Multinomial tests
  - Contingency tables (Chi-square test)
  - Difference in proportions (e.g., counts)
- Difference in Distributions
  - Histogram distances (Kullback Leibler)
  - Rank based (Wilcoxon)
  - Cumulative distribution based (Kolmogorov-Smirnov)

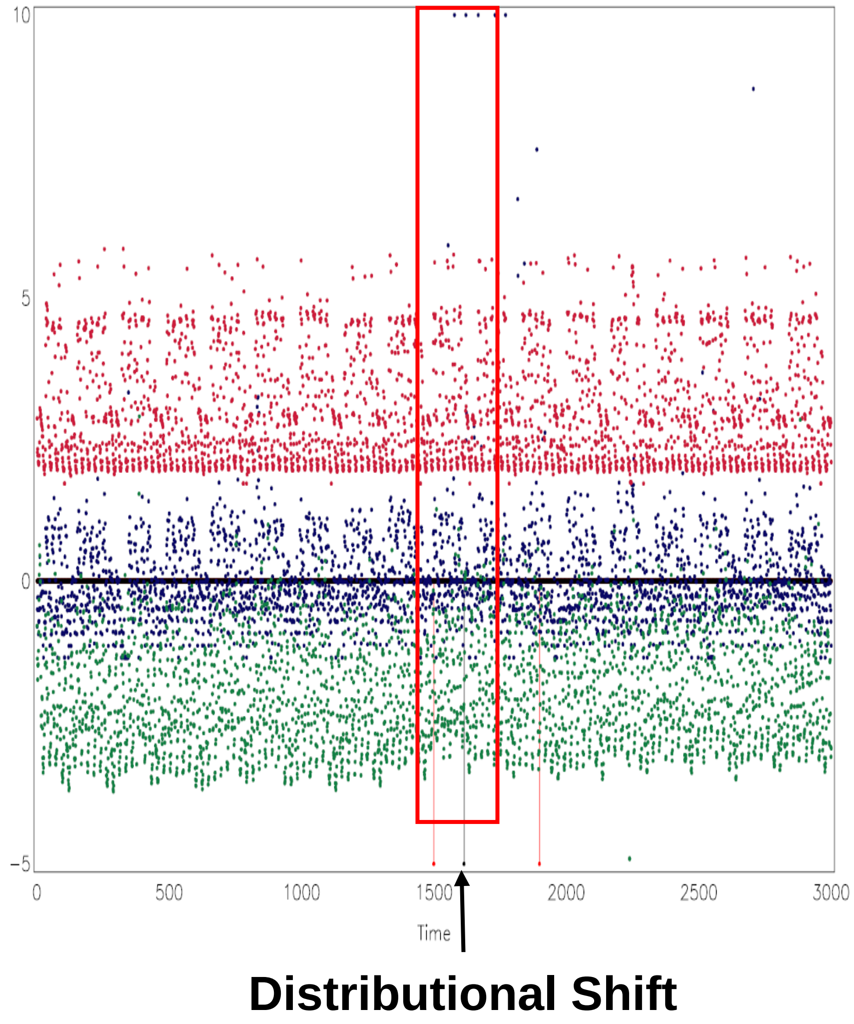
Change Detection Schemes

- *Comprehensive framework:* Detecting Changes in Data Streams. [Kifer et al., 2004]
- *Kernel based:* Statistical Change Detection in Multi-dimensional Data. [Song et al., 2007]
- *Nonparametric, fast, high-D:* Change Detection you can believe in: Finding Distributional Shifts in Data Streams. [Dasu et al., 2006, 2009]

Change (Detection) you can believe in: Finding Distributional Shifts in Data Streams

[Dasu, Krishnan, Li, Venkatasubramanian, Yi, 2009]

- Compare data distributions in two windows
  - Kdq-tree partitioning
  - Kullback-Leibler distance of histograms
    - Counts
    - Referential distance
  - Bootstrap to determine threshold
  - File descriptor data stream
    - 3 variables shown
    - Change detection led to improvement in process and cycle times



Changes in Distributions Caused by Missing/Duplicate Data

- Subtle cases of duplication/missing data
  - Result in changes in distributions
  - Missing → “lower” density regions
  - Duplicates → “higher” density regions

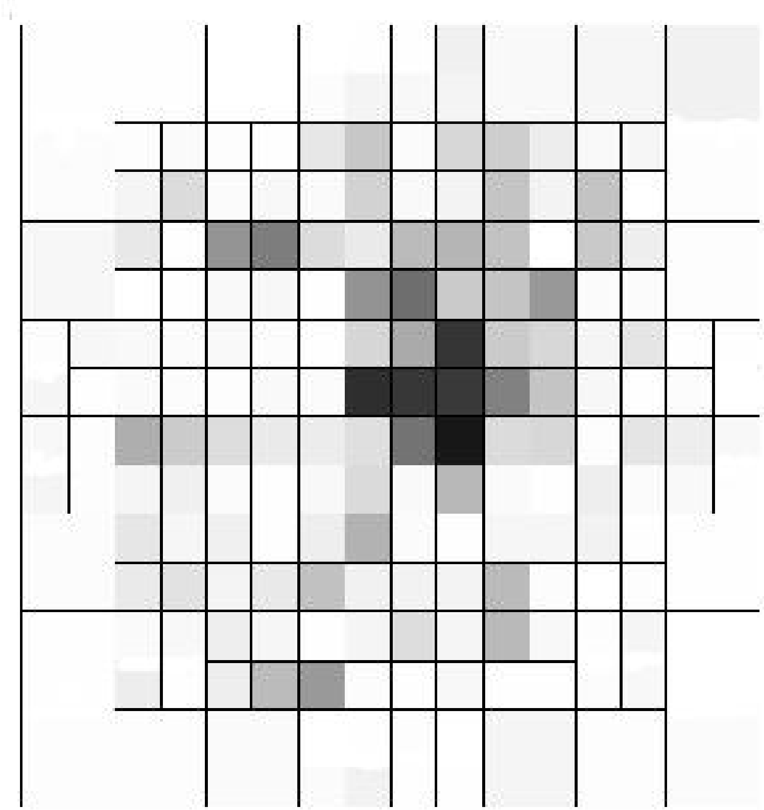
Part II. Data Quality Mining

1. Outlier Mining
2. Change Detection
3. Handling Missing and Duplicate Data

Missing Data Example

- Comparison of telecommunications data sets
- Anomalous months
  - Missing data
  - Kdq tree partition
  - Darker → greater density difference
- Automatic detection is speedy, provides an opportunity to recover and replace data before it is archived

Density change from month to another



Statistical Solutions

[Little & Rubin 1987; Allison 2002; Yuan 2000]

- Missing Value Imputation [Little & Rubin 1987; Allison 2002]
  - Point estimates
    - Mean, median
  - Model based
    - Regression
  - Simulation based
    - MCMC
  - Cautionary Tales [Allison 2000]
- Tools
  - SAS – PROCs MI and MIANALYZE
  - [Yuan 2000]

Handling Missing Data

- **Completion Using Association Rules**
  - Based on a consensus from rules with high confidence and user interaction
  - Based on measures scoring the best rules to select the replacement value [Wu et al., 2004]
- **Imputation using NN, Clustering and SVM**
  - K-Nearest Neighbour Imputation [Batista, Monard, 2003]
  - K-means Clustering Imputation [Li et al., 2004]
  - Fuzzy K-means Clustering [Acuna, Rodriguez, 2004]
  - SVM [Feng et al. 2005]

Handling Duplicate Data

[Elmagarmid et al., 2007]

Decision Model (*Prototype*)	Authors	Type
Error-based Model	[Fellegi & Sunter 1969]	Probabilistic
EM-based Method	[Dempster *et al.* 1977]	
Induction Model Clustering Model (*Tailor*)	[Bilenko et Mooney 2003] [Elfeky *et al.* 2002]	
1-1 matching	[Winkler 2004]	
Bridging File	[Winkler 2004]	
Sorted Nearest Neighbors and variants		Empirical
XML object Matching	[Weiss, Naumann 2004]	
Hierarchical Structure (*Delphi*)	[Ananthakrishna *et al.* 2002]	
Matching Prediction based on clues	[Buechi *et al.* 2003]	Knowledge-based
Instance-based functional dependencies	[Lim *et* al. 1993]	
Transformation Fuctions (*Active Atlas*)	[Tejada *et al.* 2001]	
Variant of NN based on rules for identifying and merging duplicates (*Intelliclean*)	[Low *et al.* 2001]	

Machine Learning Deduplication

Training examples

Pair	Label	f_1	f_2	...	f_n	
Customer 1, Customer 2	D	1.0	0.4	...	0.2	1
Customer 1, Customer 3	N	0.0	0.1	...	0.3	0
Customer 4, Customer 5	D	0.3	0.4	...	0.4	1

Similarity distance functions

Classifier

Learnt Rule: All-Ngrams*0.4
+ CustomerAddressNgrams*0.2
– 0.3EnrollYearDifference
+ 1.0*CustomerNameEditDist
+ 0.2*NumberOfAccountsMatch – 3 > 0

Learners:
SVMs: high accuracy with limited data [Christen, 2008]
Decision trees: interpretable, efficient to apply
Perceptrons: efficient incremental training [Bilenko et al., 2005]

Unlabeled list

Record	f_1	f_2	...	f_n	
Customer 6	0.0	0.1	...	0.3	?
Customer 7	1.0	0.4	...	0.2	?
Customer 8	0.6	0.2	...	0.5	?
Customer 9	0.7	0.1	...	0.6	?
Customer 10	0.3	0.4	...	0.4	?
Customer 11	0.0	0.1	...	0.1	?

Perspectives

Since the first *Data Quality Mining* definition:

"Deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large DBs."

[Hipp, Güntzer, Grimmer, 2001]

Recent Advances:

- Outlier mining *[Kriegel+09]*
- Change detection *[Kifer+04, Dasu+09]*
- Constraints and CFD mining *[Golab+08, Fan+09]*
- Imputation using K-means or SVM *[Li+04, Feng+05]*

Issues remain

- Treat glitches in isolation
- No connection between detection and cleaning
- No iteration of detection-cleaning
  - Cleaning introduces new glitches?
- Optimal cleaning strategies?

Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

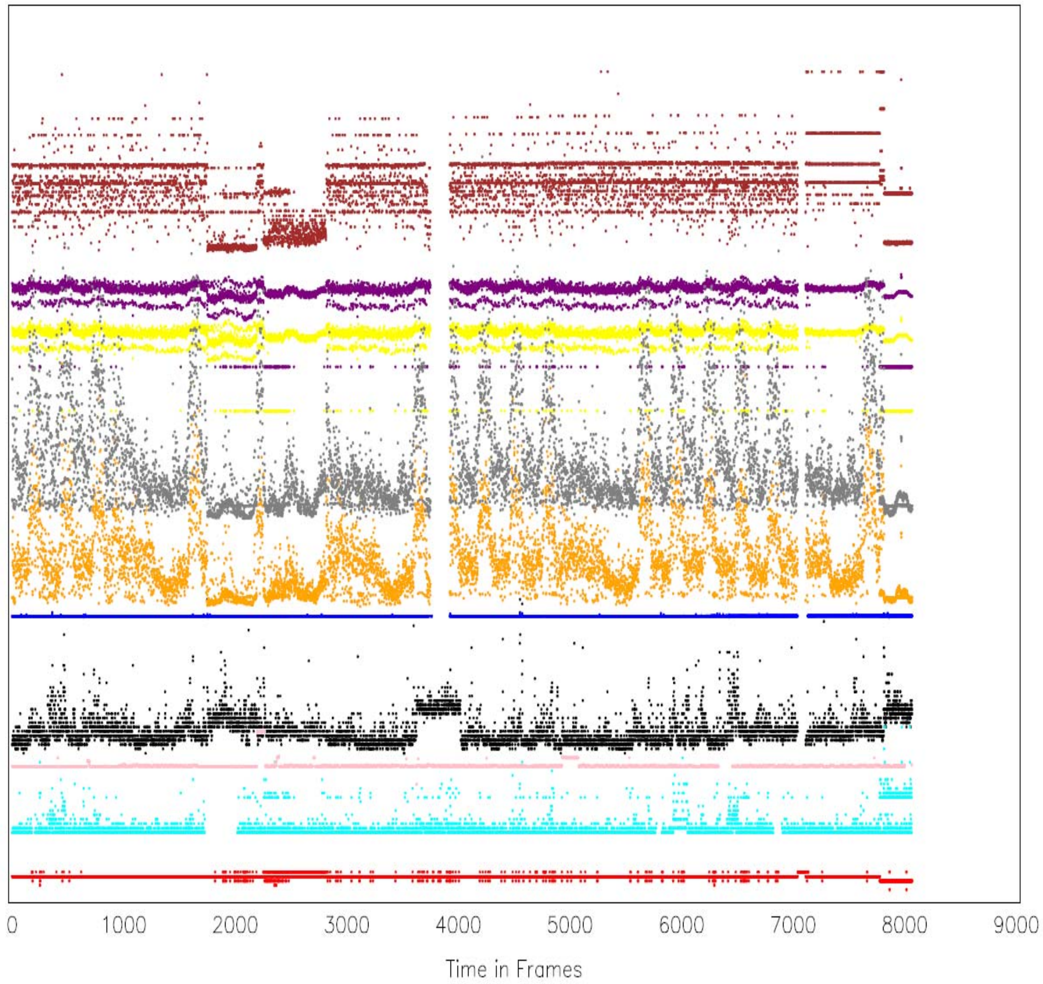
Part III. Case Study

Case Study: Networking Data

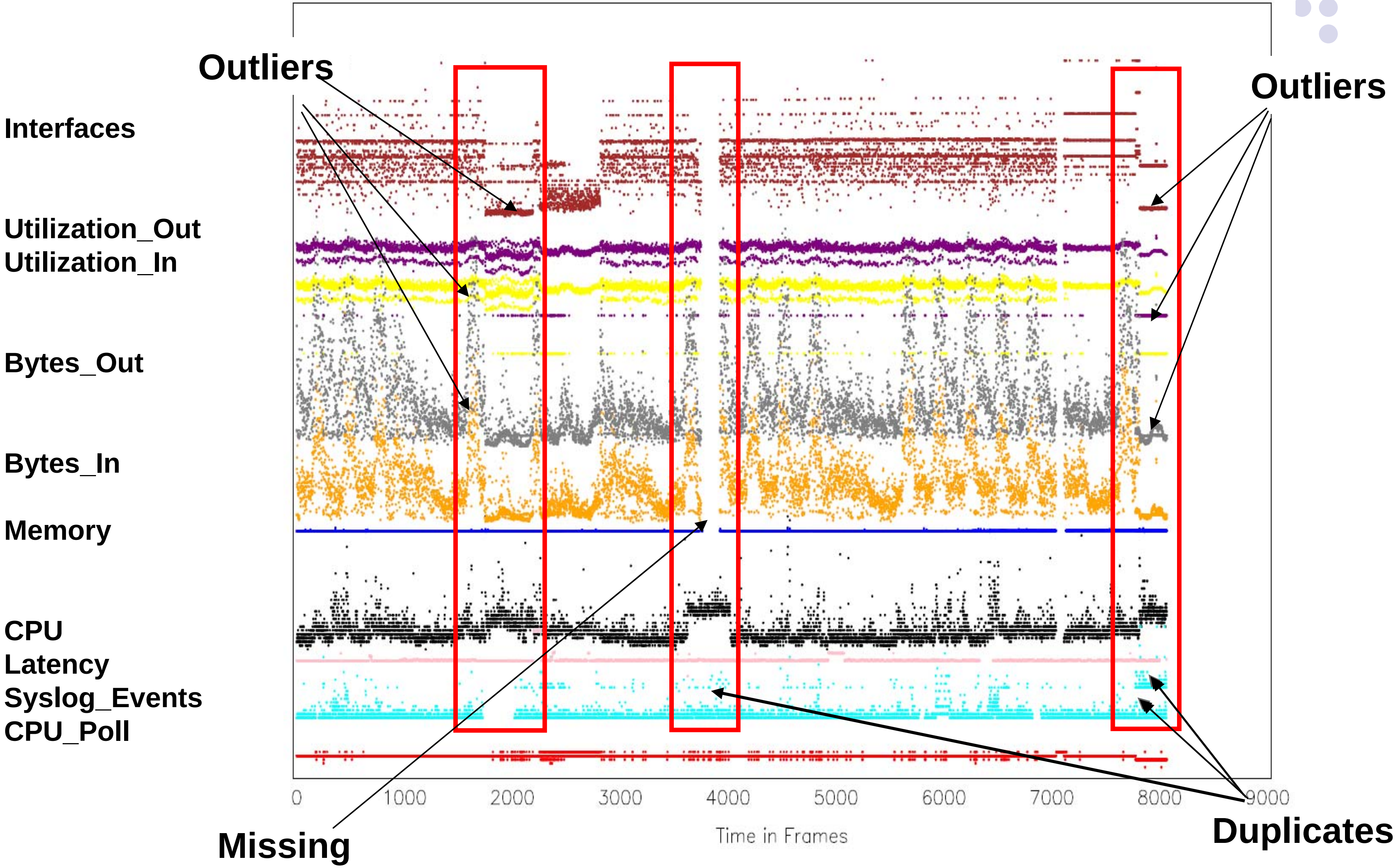
- Analyze IP data streams e.g. change detection
- Attributes
  - Resource usage
  - Traffic measurements
  - Performance metrics
  - Alarms
- Gathered from multiple, disparate sources

IP Data Streams: A Picture

- 10 Attributes, every 5 minutes, over four weeks
- Axes transformed for plotting
- Multivariate glitches!



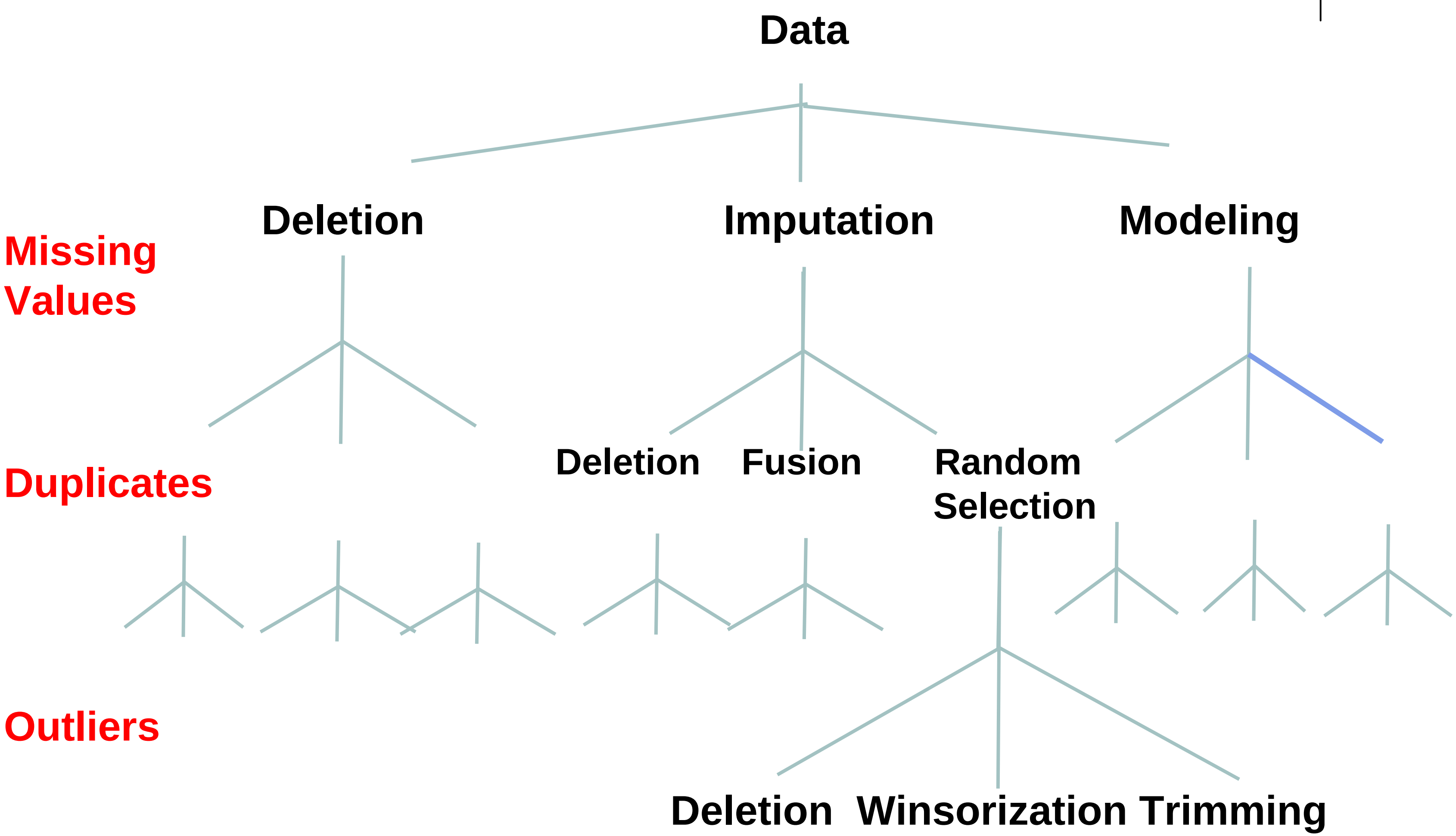
Detection of Data Glitches



What Can Be Done?

- Cleaning strategies (*ad hoc*)
  - Impute missing values → component-wise median?
  - De-duplicate → retain a random record?
  - Outliers → identify and remove? So many methods but contradicting results?
  - Drop all records that have <u>any</u> imperfection
  - Add special categories and analyze singularities in isolation
- Almost all existing approaches look at one-shot approaches to univariate glitches. Why?

So Many Choices …



How do we reduce the number of choices? Which one is the best?

Ordering in cleaning matters

Input: Dirty dataset

poll#	date	time	inrate	outrate
1	08/31/2009	00:56:00	130	130
~~2~~	~~09/01/2009~~	~~00:01:00~~	~~120~~	~~130~~
3	09/01/2009	00:01:00	130	150
4	09/01/2009	00:05:01	130	130
~~5~~	~~09/01/2009~~	~~00:10:01~~	~~110~~	~~100~~
~~6~~	~~09/01/2009~~	~~00:10:01~~	~~140~~	~~140~~
7	09/01/2009	00:10:01	130	110

Cleaning

Alternatives:

Replacement Only

median	130	130
sum	630	630

1 Replacement 2 Fusion

median	130	130
sum	520	520

Ordering in cleaning matters

Input: Dirty dataset

poll#	date	time	inrate	outrate
1	08/31/2009	00:56:00	-	-
2	09/01/2009	00:01:00	120	130
3	09/01/2009	00:01:00	130	150
4	09/01/2009	00:05:01	-	-
5	09/01/2009	00:10:01	110	100
6	09/01/2009	00:10:01	140	140
7	09/01/2009	00:10:01	130	110

Cleaning

Alternatives:

1 Fusion 2 Replacement

Fusion choices impact replacement and may mask/generate inconsistencies

	#2 #5		#2 #6		#2 #7		#3 #5		#3 #6		#3 #7	
median	115	115	130	135	125	120	120	125	135	145	130	130
sum-before	230	230	260	270	250	240	240	250	270	290	260	260

Constraint: $\Sigma \text{inrate} = \Sigma \text{outrate}$

A Cleaning Strategy Based On Explainable Patterns

Input:

Dirty dataset

Cleaning

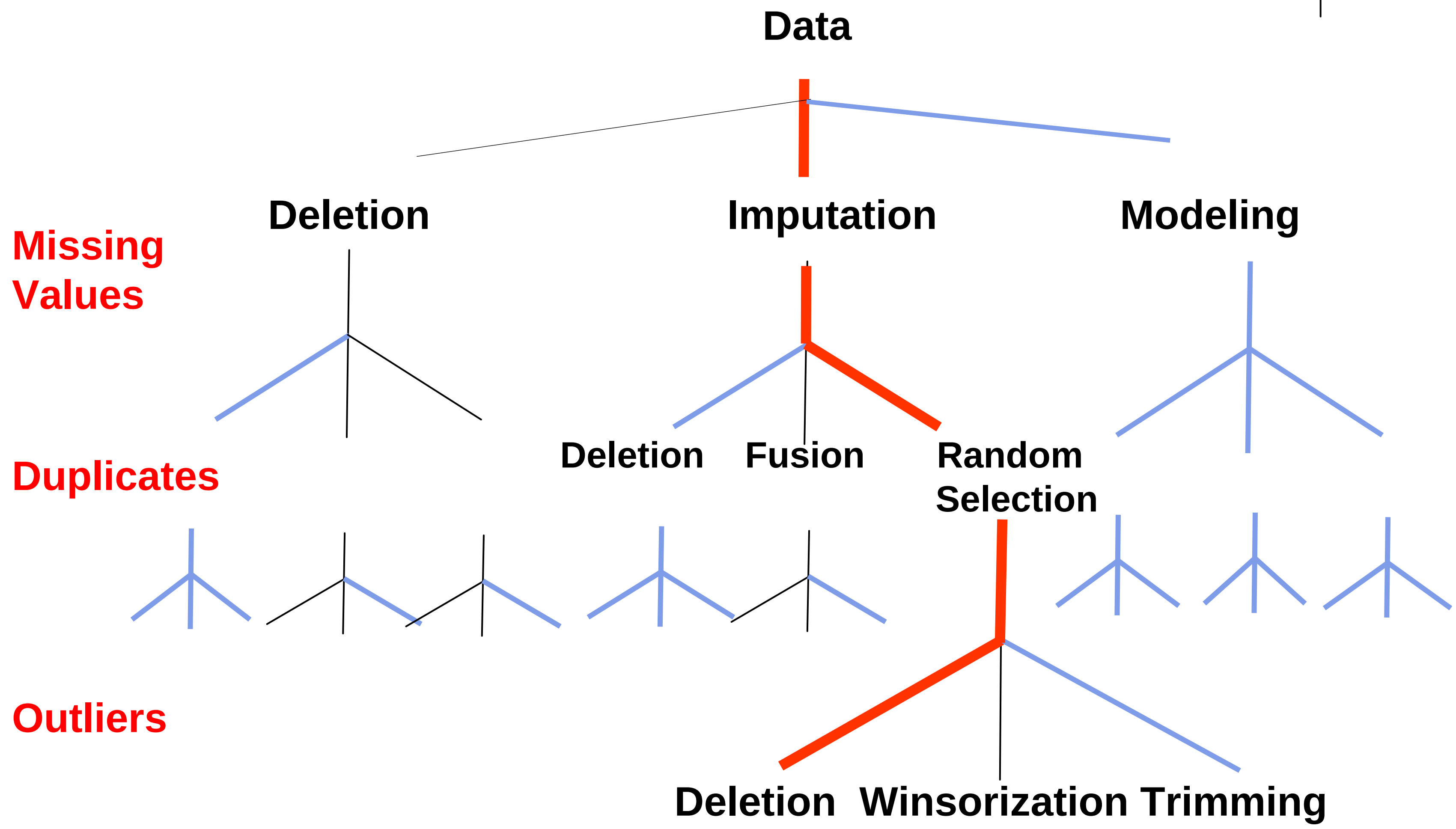
poll#	date	time	inrate	outrate
1	08/31/2009	00:56:00	missing	missing
2	09/01/2009	00:01:00	duplicate	
3	09/01/2009	00:01:00	duplicate	
4	09/01/2009	00:05:01	missing	missing
5	09/01/2009	00:10:01	duplicate	
6	09/01/2009	00:10:01	duplicate	
7	09/01/2009	00:10:01	duplicate	

Explanation

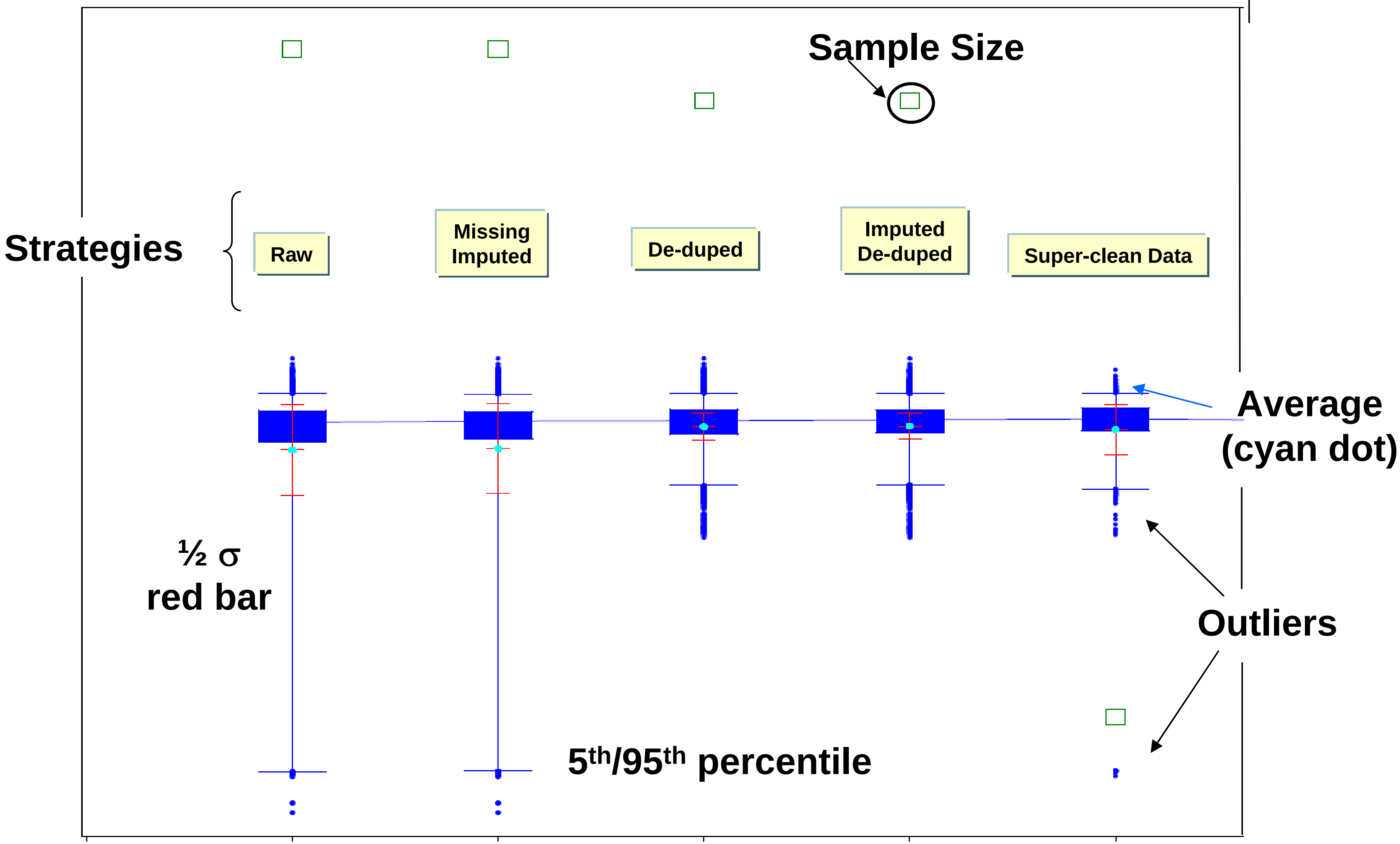
Replacement

Using the values of the first adjacent duplicates

A particular path → A sequence of strategies



Cleaning Strategies: Boxplot Comparison



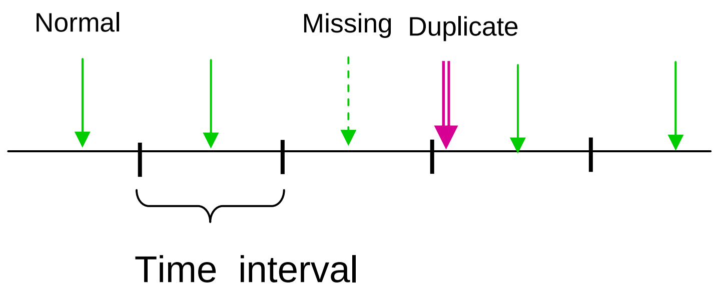
Can We Do Better?

- We used no domain knowledge or any data-specific property
- Are there any patterns in the glitches that we can exploit to develop powerful cleaning strategies?
- Can we provide any statistical guarantees on the "clean" data sets? A statistical notion of "best"?

What Do We Mean By Patterns of Glitches?

- Univariate/Multivariate Combination of DQ problems
  - Complex patterns (co-occurring & lagged)
  - outliers and missing values
  - outliers and duplicates
  - **missing and duplicates**

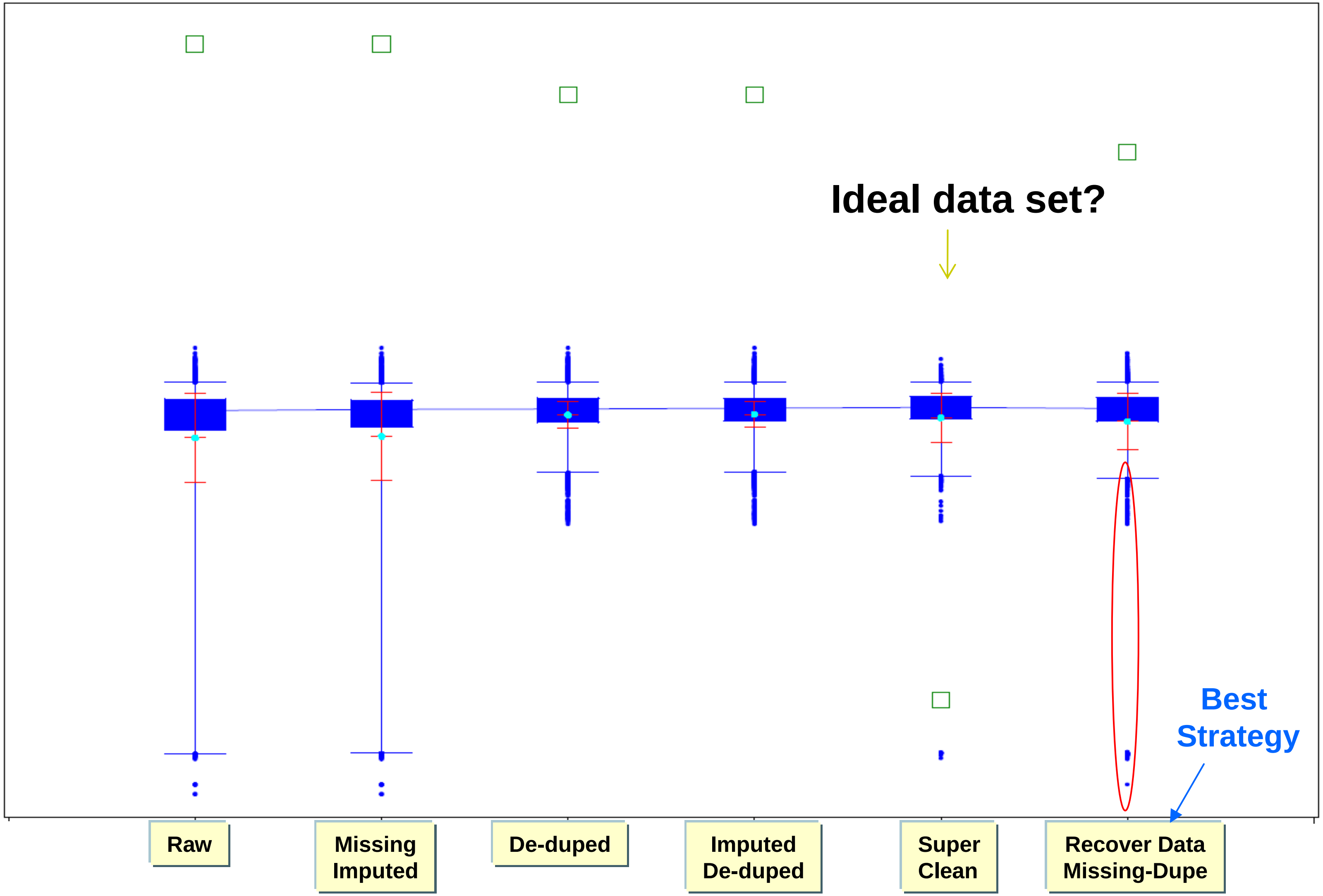
Missing-Duplicate pairs



Good News:

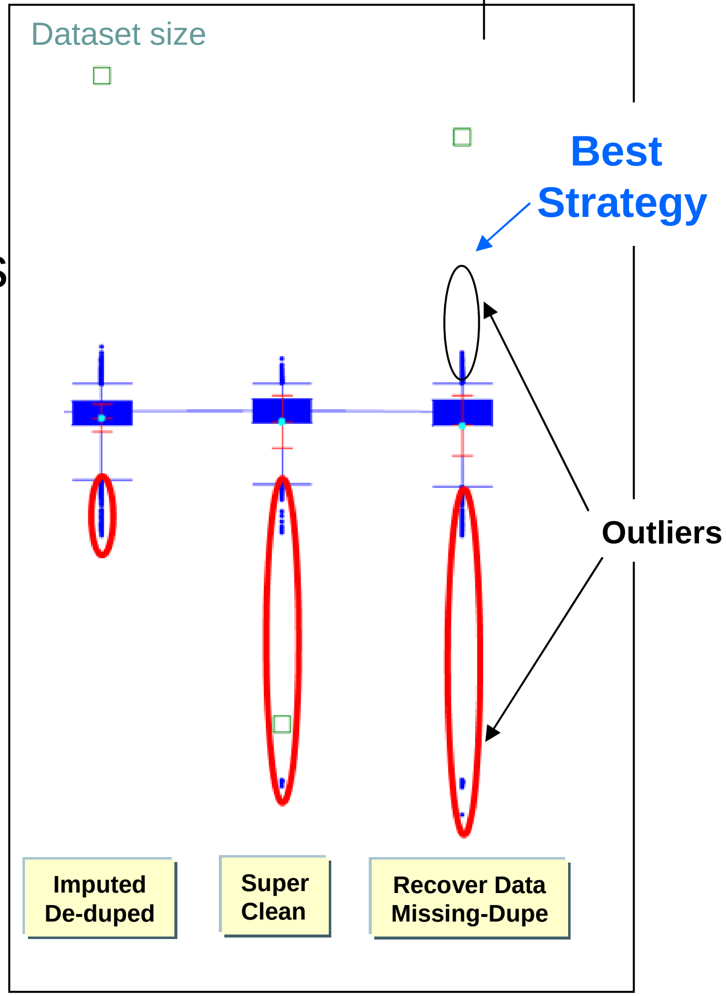
- Artifact of collecting mechanism
- Drive our cleaning strategy!

How To Select the Best Cleaning Strategy?



It depends on what matters most…

- Two alternatives for cleaning:
  - Discovered patterns and domain knowledge-driven replacement of missing values with adjacent duplicates
  - Quantitative cleaning, e.g., blind imputation
- Note
  - Blind imputation misses outliers
  - Additional iterations are needed because cleaning reveals new glitches



Case Study: Conclusion

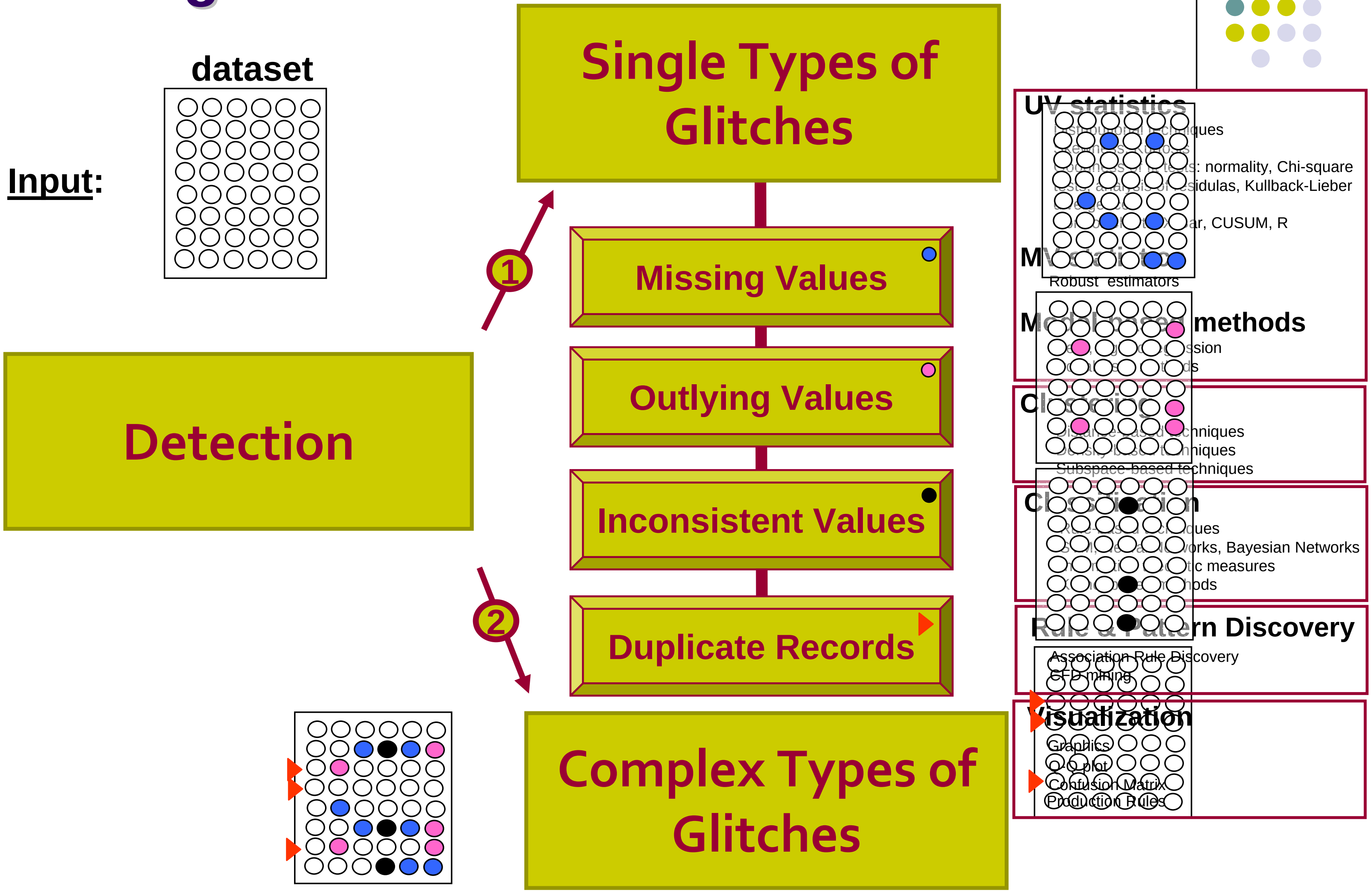
- IP data stream – multivariate, massive, glitchy
- Critical for network monitoring
- Patterns and dependencies in glitches are used to recover much of the data such that the treated dataset is close to the ideal dataset
- Discovery of explanatory variables is useful for understanding recurrent DQ problems

New Research Directions

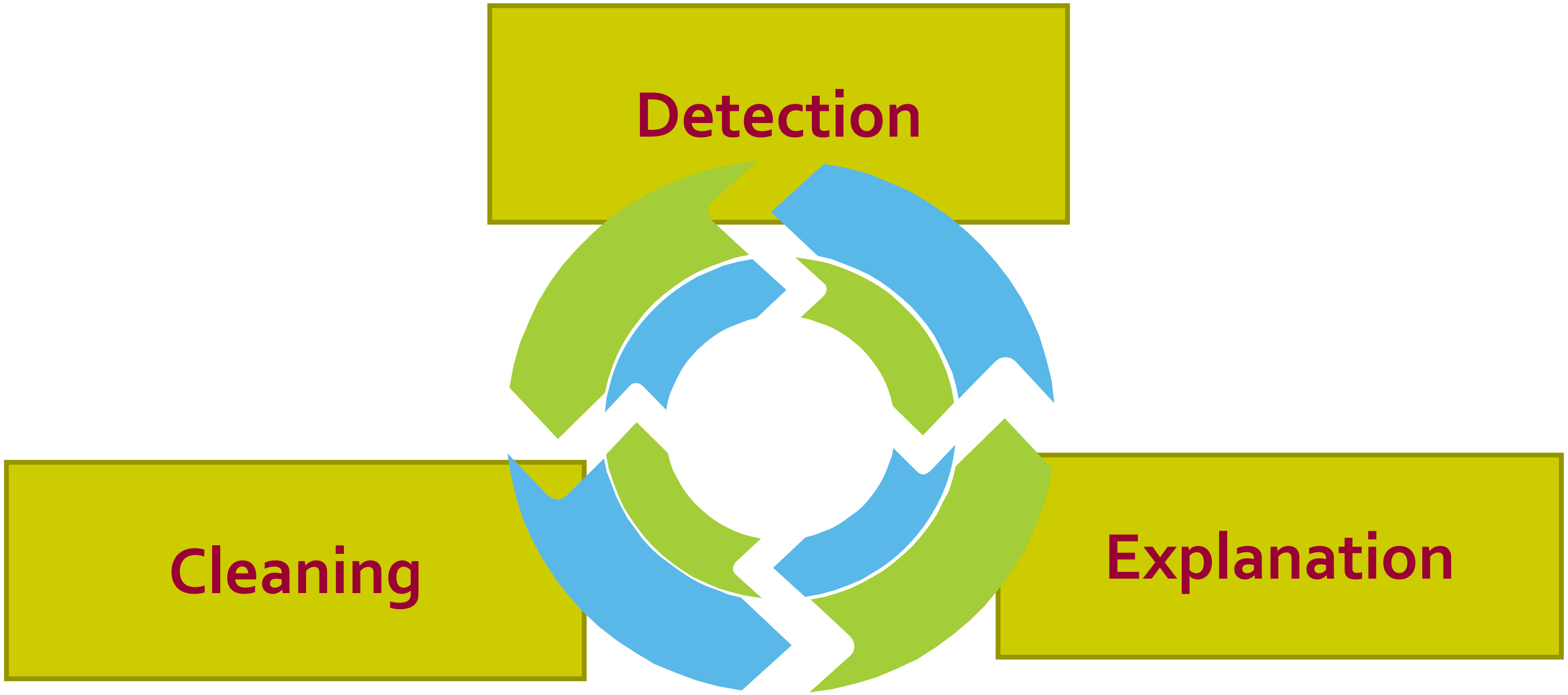
Discovering complex and concomitant data glitches

- Single → Complex, multivariate glitches
- Connecting detection with cleaning
  - Iteration
  - Explanation
- Identifying candidate strategies and choosing the best strategy

New Direction 1:
Model glitches as multivariate entities

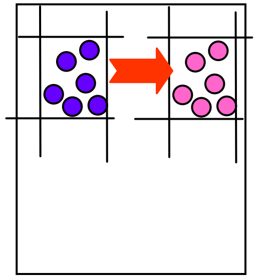


New Direction 2 : Connect detection and cleaning



New Direction 2 : Connect detection and cleaning

Detection → **Explanation**



Glitch set characteristics:

- Distribution
- Locality
- Density
- Variety of glitches
- Commonality (shared conditions on the dataset)
- Relationships and correlations
- Dynamics (common trends, concomitance)

Clues for the root causes:

- Localization of deficiencies in a data source/process
- Severity of the deficiencies
- Nature and extent of the deficiencies
- Specificity of deficiencies
- Propagation mechanisms
- Punctual/recurrent error generation

New Direction 3: Select best cleaning strategy

Many choices: automation & repeatability required

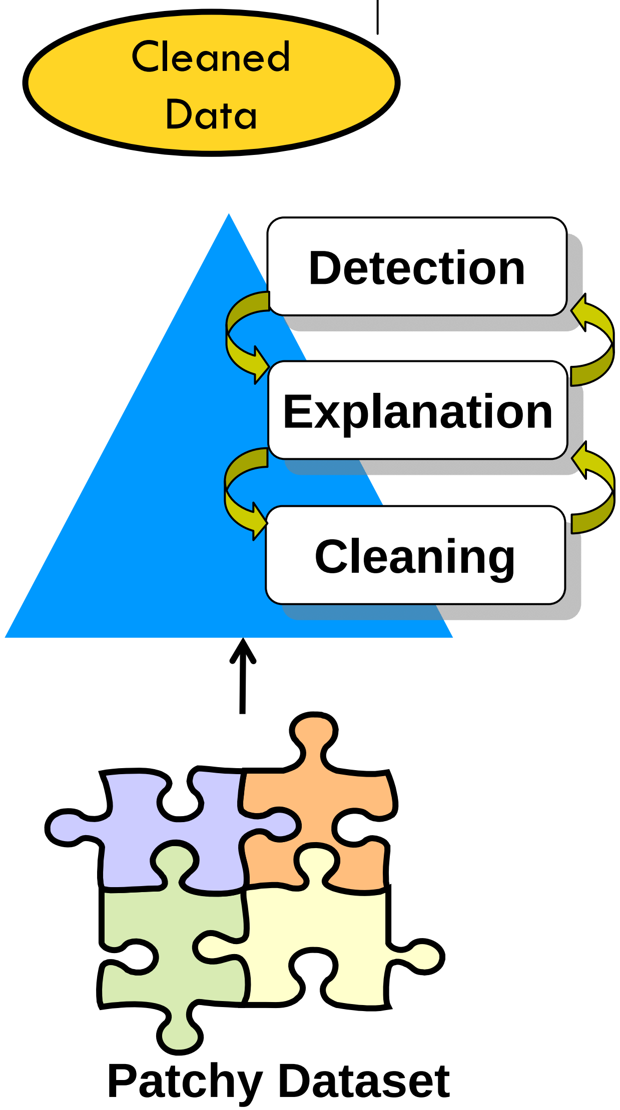
- Identify candidate strategies
  - Cost
  - Glitch reduction
- Select the best strategy
  - Distance from original
  - Distance from ideal

Interesting research questions

- **Glitch scoring**
  - Conflict resolution: multiple methods, same glitch type
  - Weighting, combining scores: multiple glitch types, same value
  - Choosing threshold values: same pattern, multiple thresholds?
- **Patterns of glitches: significance**
  - Test of independence of glitches?
  - Spatio-temporal patterns?
- **Bias**
  - Impact of mutual masking effect, order of treatments

Overview

- Data Quality Research
- Advanced techniques in DQM
- Motivating Case study
- New Directions for DQM



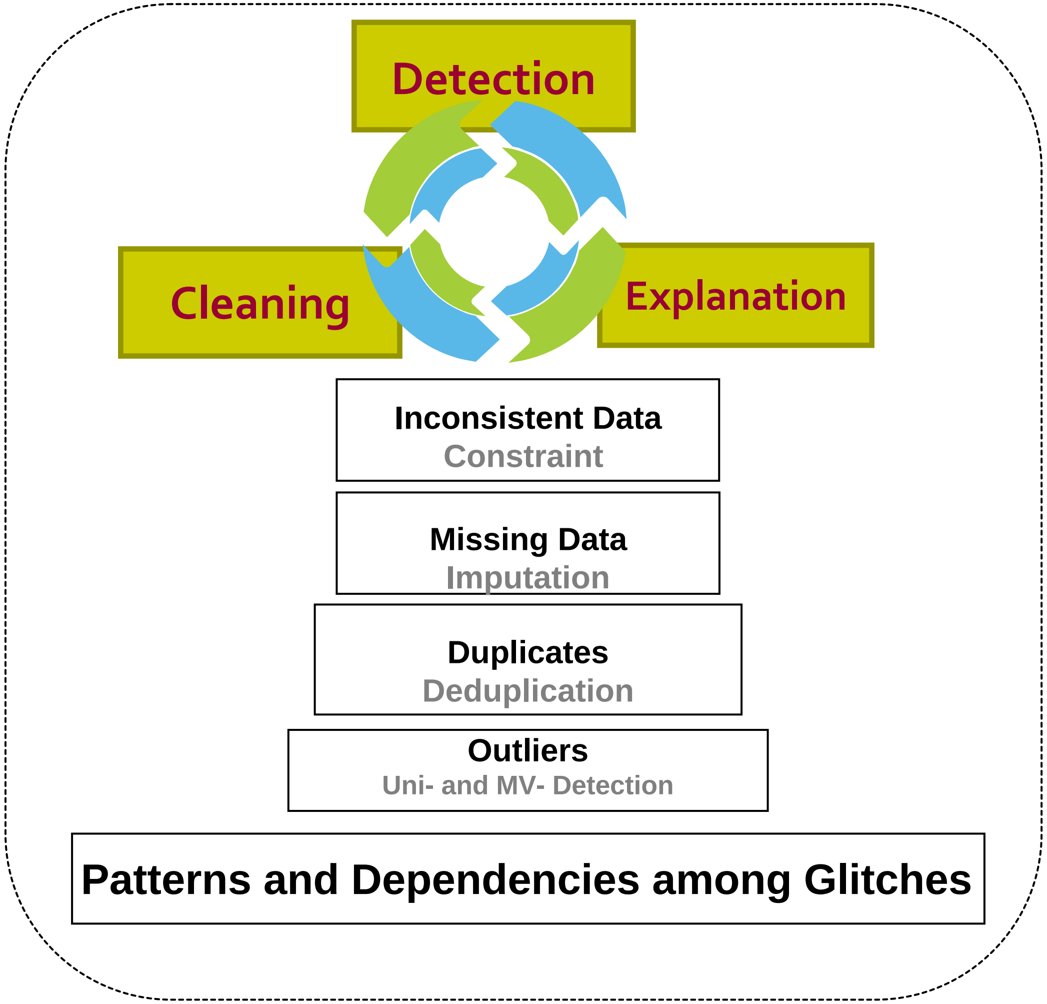
DQM Summary: Multivariate Glitches

- Glitches are multivariate with strong interdependencies
  - Static & temporal
  - Domain and application dependent
- DQM framework is important
  - Extant approaches tend to treat each class of glitches separately – misleading.
- Patterns and distribution of glitches are crucial in formulating cleaning strategies

DQM Summary: Process and Strategies

- Iterative and complementary cleaning strategies
- Best DQM strategies
  - Quantitative criteria
  - Resource-dependent
  - Domain, user and operational needs

Iterative Detection and Cleaning



Conclusion

DQM Challenges

- Dimensionality and complexity
- Uncertainty and ambiguity
- Dynamic nature
- Benchmarking

DQM Framework

- Multidisciplinary approach
- Unified process
- Repeatability
- Statistical guarantees

Thanks

Any questions?

Limited Bibliography

References

Books

- BATINI, Carlo, SCANNAPIECO, Monica. Data Quality Concepts, Methodologies and Techniques. Data-Centric Systems and Applications. Springer-Verlag, 2006.
- BARNETT, V., LEWIS, T., Outliers in statistical data. John Wiley, Chichester, 1994.
- DASU, Tamraparni, JOHNSON, Theodore. Exploratory Data Mining and Data Cleaning. John Wiley, 2003.
- HAWKINS, D., Identification of Outliers. Chapman and Hall, London, 1980.
- HERZOG, Thomas N., SCHEUREN, Fritz J., WINKLER, William E., Data Quality and Record Linkage Techniques, Springer, May 2007.
- KIMBALL, Ralph, CASERTA, Joe. The Data Warehouse ETL Toolkit, Wiley, 2004.
- NAUMANN, Felix Quality-Driven Query Answering for Integrated Information Systems. Lecture Notes in Computer Science, vol. 2261. Springer-Verlag, 2002.
- Tukey, John Wilder. Exploratory Data Analysis. Addison-Wesley, 1977
- WANG, Richard Y., ZIAD, Mostapha, LEE, Yang W. Data Quality. Advances in Database Systems, vol. 23. Kluwer Academic Publishers, 2002.

Surveys

- CHANDOLA, Varun, BANERJEE, Arindam, KUMAR, Vipin, Anomaly Detection A Survey. ACM Computing Surveys, September 2009.
- ELMAGARMID, Ahmed K., IPEIROTIS, Panagiotis G., VERYKIOS, Vassilios S., Duplicate Record Detection A Survey, IEEE Transations on knowledge and Data Engineering (TKDE) Vol. 19 No. 1 January 2007, pp. 1-16.
- HELLERSTEIN, Joseph, Quantitative Data Cleaning for Large Databases. White paper, United Nations Economic Commission for Europe, February, 2008. http://db.cs.berkeley.edu/jmh/cleaning-unece.pdf
- NAVARRO, Gonzalo. A Guided Tour to Approximate String Matching. ACM Comput. Surv., 33(1), pp. 31–88, 2001.
- WINKLER, William E., Overview of Record Linkage and Current Research Directions, Tech. Rep. of U.S. Census Bureau, February. 2006 http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

References

Tutorials

- BATINI, Carlo, CATARCI, Tiziana, SCANNAPIECO, Monica. A Survey of Data Quality Issues in Cooperative Systems. Tutorial ER 2004.
- KOUDAS, Nick, SARAWAGI, Sunita, SRIVASTAVA, Divesh. Record Linkage Similarity Measures and Algorithms. Tutorial SIGMOD 2006.
- BANERJEE, Arindam, CHANDOLA, Varun, KUMAR, Vipin, SRIVASTAVA Jaideep, LAZAREVIC, Aleksandar. Anomaly Detection A Tutorial. Tutorial SIAM Conf. on Data Mining 2008.
- KRIEGEL, Hans-Peter, KROGER, Peer, ZIMEK, Arthur. Outlier Detection Techniques. Tutorial, PAKDD 2009. http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/tutorial_slides.pdf

Data Profiling

- CARUSO, FRANCESCO, COCHINWALA, MUNIR, GANAPATHY, UMA, LALK, GAIL, MISSIER, PAOLO. 2000. Telcordia’s Database Reconciliation and Data Quality Analysis Tool. Proc. VLDB 2000, pp. 615–618, 2000.
- DASU, TAMRAPARNI, JOHNSON, THEODORE, S. Muthukrishnan, V. Shkapenyuk, Mining Database Structure; Or, How to Build a Data Quality Browser, Proc. SIGMOD 2002.

Data Preparation and Data Quality Mining

- HIPP, J., GUNTZER, U., GRIMMER, U. Data Quality Mining - Making a Virtue of Necessity. Proc. Workshop DMKD 2001.
- LUBBERS, D., GRIMMER, U., JARKE, M. Systematic Development of Data Mining-Based Data Quality Tools. Proc. VLDB 2003, pp. 548-559, 2003.
- KLINE, R.B., Data Preparation and Screening, Chapter 3. in Principles and Practice of Structural Equation Modeling, NY Guilford Press, pp. 45-62, 2005.
- PEARSON, Ronald K. Surveying Data for Patchy Structure. SDM 2005.
- STATNOTES Topics in Multivariate Analysis. Retrieved 10/17/2008 from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

References

Data Cleaning – ETL

- BILKE, Alexander, BEIHOLDER, Jens, BOHM, Christoph, DRABA Karsten, NAUMANN, Felix, WEIS, Melanie. Automatic Data Fusion with HumMer. Proc. VLDB 2005 1251-1254, 2005.
- CHAUDHURI, Surajit, GANTI, Venkateh, KAUSHIK, Raghav. A Primitive Operator for Similarity Joins in Data Cleaning. Proc. ICDE 2006.
- CHRISTEN, Peter. Febrl an open source data cleaning, deduplication and record linkage system with a graphical user interface. KDD 2008, pp. 1065-1068, 2008.
- CHRISTEN, Peter, CHURCHES, Tim, ZHU, Xi. Probabilistic name and address cleaning and standardization. Proc. Australasian Data Mining Workshop 2002. http://cs.anu.edu.au/~Peter.Christen/publications/adm2002-cleaning.pdf
- GALHARDAS, Helena, FLORESCU, Daniela, SHASHA, Dennis, SIMON, Eric, SAITA, Cristian-Augustin. Declarative Data Cleaning Language, Model, and Algorithms, Proc. VLDB Conf., pp. 371-380, 2001.
- HERNANDEZ, M., STOLFO, S., Real-World Data is Dirty Data Cleansing and the Merge/Purge Problem, Data Mining and Knowledge Discovery, 2(1)9-37, 1998.
- RAHM, E., DO, H.H., Data Cleaning Problems and Current Approaches, Data Engineering Bulletin, 23(4) 3-13, 2000.
- RAMAN, V., HELLERSTEIN, J.M. Potter's Wheel: An Interactive Data Cleaning System. Proc. VLDB 2001, pp. 381-390, 2001.
- VASSILIADIS, P., VAGENA, Z., SKIADOPOULOS, S., KARAYANNIDIS, N., SELLIS, T. ARKTOS A Tool For Data Cleaning and Transformation in Data Warehouse Environments. Bulletin of the Technical Committee on Data Engineering, 23(4), pp. 42-47, 2000.
- VASSILIADIS, P., KARAGIANNIS A., TZIOVARA, V., SIMITSIS, A. Towards a Benchmark for ETL Workflows. Proc. QDB 2007, pp. 49-60, 2007.
- WEIS, Melanie, MANOLESCU, Ioana. XClean in Action (Demo). CIDR 2007, pp. 259-262, 2007.

References

Record Linkage and duplicate detection (1/2)

- ANANTHAKRISHNA, ROHIT, CHAUDHURI, SURAJIT, GANTI, VENKATESH. Eliminating Fuzzy Duplicates in Data Warehouses. pp. 586–597, Proc. of VLDB 2002.
- BANSAL, NIKHIL, BLUM, AVRIM, CHAWLA, SHUCHI. Correlation clustering. Machine Learning, 56(1-3):89–113, 2004.
- BAXTER, ROHAN A., CHRISTEN, PETER, CHURCHES, TIM. A Comparison of Fast Blocking Methods for Record Linkage. pp. 27–29 Proc. of the KDD’03 Workshop on Data Cleaning, Record Linkage and Object Consolidation, 2003.
- BHATTACHARYA, INDRAJIT, GETOOR, LISE. Iterative Record Linkage for Cleaning and Integration. pp. 11–18 Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD, 2004.
- BHATTACHARYA, INDRAJIT, GETOOR, LISE. Collective entity resolution in relational data. TKDD, 1(1), 2007.
- BILENKO, MIKHAIL, MOONEY, RAYMOND J. Adaptive Duplicate Detection Using Learnable String Similarity Measures. Proc. KDD 2003, pp. 39–48, 2003.
- BILENKO, MIKHAIL, BASU, SUGATO, SAHAMI, MEHRAN. 2005. Adaptive Product Normalization Using Online Learning for Record Linkage in Comparison Shopping. Proc. ICDM 2005, pp. 58–65, 2005.
- CHRISTEN, Peter, Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification, ACM SIGKDD 2008 Conf., Las Vegas, August 2008.
- ELFEKY, MOHAMED G., ELMAGARMID, AHMED K., VERYKIOS, VASSILIOS S. TAILOR A Record Linkage Tool Box. pp. 17–28 Proc. of the 18th International Conf. on Data Engineering, ICDE 2002. San Jose, CA, USA, 2002.
- ELMAGARMID, AHMED K., IPEIROTIS, PANAGIOTIS G., VERYKIOS, VASSILIOS S. Duplicate Record Detection A Survey. IEEE Trans. Know. Data Eng., 19(1), 1–16, 2007.
- FELLEGI, IVAN P., SUNTER, A.B. A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183–1210, 1969.

References

Record Linkage and duplicate detection (2/2)

- GRAVANO, Luis, IPEIROTIS, Panagiotis G., JAGADISH, H. V., KOUDAS, Nick, MUTHUKRISHNAN, S., PIETARINEN, Lauri, SRIVASTAVA, Divesh. Using q-grams in a DBMS for Approximate String Processing. IEEE Data Eng. Bull., 24(4), 28–34, 2001.
- GRAVANO, LUIS, IPEIROTIS, PANAGIOTIS G., KOUDAS, NICK, SRIVASTAVA, DIVESH. Text Joins for Data Cleansing and Integration in an RDBMS. Proc. ICDE 2003, pp. 729–731, Bangalore, India, 2003.
- HERNANDEZ, M., STOLFO, S., The Merge/Purge Problem for Large Databases, Proc. SIGMOD Conf pg 127-135, 1995.
- LOW, WAI LUP, LEE, MONG-LI, LING, TOK WANG. A Knowledge-Based Approach for Duplicate Elimination in Data Cleaning. Inf. Syst., 26(8), 585–606, 2001.
- KANG, Hyunmo, GETOOR, Lise, SHNEIDERMAN, Ben, BILGIC, Mustafa, LICAMELE, Louis. Interactive Entity Resolution in Relational Data: A Visual Analytic Tool and Its Evaluation. IEEE Trans. Vis. Comput. Graph. 14(5), pp. 999-1014, 2008.
- MCCALLUM, ANDREW, NIGAM, KAMAL, UNGAR, LYLE H. 2000. Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. Proc. KDD 2000, pp. 169–178. Boston, MA, USA.
- MONGE, ALVARO E. 2000. Matching Algorithms within a Duplicate Detection System. IEEE Data Eng. Bull., 23(4), 14–20.
- TEJADA, SHEILA, KNOBLOCK, CRAIG A., MINTON, STEVEN. 2002. Learning Domain-Independent String Transformation Weights for High Accuracy Object Identification. Proc. KDD 2002, pp. 350–359, 2002.
- WEIS, MELANIE, NAUMANN, FELIX, BROSY, FRANZISKA. 2006. A Duplicate Detection Benchmark for XML (and Relational) Data. Proc. ACM SIGMOD 2006 Workshop on Information Quality in Information Systems, IQIS 2006. Chicago, IL, USA.
- WINKLER, WILLIAM E. Methods for Evaluating and Creating Data Quality. Inf. Syst., 29(7), 531–550, 2004.
- WINKLER, WILLIAM E., THIBAUDEAU, YVES. An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census. Tech. Rept. Statistical Research Report Series RR91/09. U.S. Bureau of the Census, Washington, DC, USA, 1991.

References

Inconsistencies

- BOHANNON, Philip, FAN Wenfei, GEERTS, Floris, JIA, Xibei, KEMENTSIETSIDIS, Anastasios Conditional Functional Dependencies for Data Cleaning. Proc. ICDE 2007, pp. 746-755.
- BRAVO, Loreto, FAN, Wenfei, MA, Shuai. Extending Dependencies with Conditions. Proc. VLDB 2007, pp. 243-254.
- CERI, Stefano, Di GIUNTA, Francesco, LANZI, Pier Luca. Mining constraint violations. ACM Trans. Database Syst., 32(1): 6, 2007.
- CHANDEL, A., KOUDAS, Nick, PU, K., SRIVASTAVA Divesh. Fast Identication of Relational Constraint Violations. Proc. ICDE 2007.
- FAN, Wenfei, GEERTS, Floris, KEMENTSIETSIDIS, Anastasios Conditional functional dependencies for capturing data inconsistencies. TODS:33(2), June 2008.
- FAN, Wenfei, GEERTS, Floris, JIA, Xibei Semandaq A Data Quality System Based on Conditional Functional Dependencies, VLDB'08, (demo), 2008.
- FAN, Wenfei, GEERTS, Floris, LAKSHMANAN, Laks V. S., XIONG, Ming. Discovering Conditional Functional Dependencies. Proc. ICDE 2009, pp. 1231-1234.
- GOLAB, Lukasz, KARLOFF, Howard J., KORN, Flip, SRIVASTAVA Divesh, YU, Bei. On generating near-optimal tableaux for conditional functional dependencies. PVLDB 1(1) 376-390, 2008.
- KORN, Flip, MUTHUKRISHNAN S., ZHU, Yunyue Checks and Balances Monitoring Data Quality Problems in Network Traffic Databases. Proc. VLDB 2003, pp. 536-547.

Change Detection

- AGGARWAL, C. C. A framework for diagnosing changes in evolving data streams. Proc. ACM SIGMOD 2003.
- DASU, T., KRISHNAN S., LIN, D., VENKATASUBRAMANIAN, S., YI, K. Change (Detection) you can believe in Finding Distributional Shifts in Data streams. Proc. IDA'09, 2009.
- DASU, T., KRISHNAN S., VENKATASUBRAMANIAN, S., YI, K. An information-theoretic approach to detecting changes in multi-dimensional data streams. Proc. Interface'06, 2006.
- SONG, X., WU, M., JERMAINE, C., RANKA S. Statistical change detection for multidimensional data. Proc. ACM SIGKDD'07, pp. 667-676, 2007.

References

Outlier Detection (1/2)

- AGARWAL, D., Detecting anomalies in cross-classified streams a Bayesian approach. Know. Inf. Syst., 11(1):29–44, 2006.
- ANGIULLI, F., PRIZZUTI, C., Fast Outlier Detection in High Dimensional Spaces. Proc. Conf. on Principles of Data Mining and Knowledge Discovery, pp. 15–26, 2002.
- BAY, D.S., SCHWABACHER, M., Mining distance-based outliers in near linear time with randomization and a simple pruning rule. Proc. KDD 2003.
- BREUNIG, M., KRIEGEL, H-P., NG, R.T., SANDER, J., LOF Identifying Density-Based Local Outliers. Proc. of the 2000 ACM SIGMOD International Conf. on Management of Data, pp. 93–104. Dallas, TX, USA, 2000.
- CHEN, Y., DANG, X., PENG, H., and BART, H., Outlier detection with the kernelized spatial depth function. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008.
- CORMODE, G., HADJIELEFTHERIOU, M., Finding frequent items in data streams. Proc. VLDB 2008.
- ESKIN, E., Anomaly detection over noisy data using learned probability distributions. Proc. ICML 2000, pp. 255-262, 2000.
- FILZMOSER, P., MARONNA, R., WERNER, M. Outlier detection in high dimensions. Computational Statistics and Data Analysis, 52, pp. 1694–1711, 2008.
- GALEANO, P., PENA, D., TSAY, R. S. Outlier detection in multivariate time series by projection pursuit. Journal of American Statistical Association, 101(474):654–669, 2006.
- HAN, F., WANG, Y., WANG H., Odabk: An effective approach to detecting outlier in data stream. Proc. Intl. Conf. on Mach. Learn. and Cybernetics, pp. 1036–1041, 2006.
- HE, Z., XU, X., DENG, S., Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10), pp. 1641–1650, 2003.
- HUBERT, M., VADER VEEKEN, S., Outlier detection for skewed data. Journal of Chemometrics, 22, pp. 235–246, 2007.
- JIANG, S.-Y., LI, Q.-H., LI, K.-L., WANG, H., MENG, Z.-L., GLOF a new approach for mining local outlier. Proc. Int. Conf. Mach. Learn. Cybernetics, vol. 11, pp. 157–162, 2003. JIN, W., TUNG, A.K.H., HAN, J., Mining Top-n Local Outliers in Large Databases. Proc. KDD 2001, pp. 293–298, 2001.
- KIFER, D., BEN-DAVID, S., GEHRKE, J., Detecting changes in data streams. Proc. VLDB 2004, pages 180–191, 2004.

References

Outlier Detection (2/2)

- KNORR, Edwin M., NG, Raymond T., Algorithms for Mining Distance-Based Outliers in Large Datasets. Proc. VLDB 1998, pp. 392–403, 1998.
- LIU, R., SINGH, K., TENG, J., Ddma-charts: Nonparametric multivariate moving average control charts based on data depth. Advances in Statistical Analysis, 88, pp. 235–258, 2004.
- KRIEGEL, H.-P., SCHUBERT, M., ZIMEK, A., Angle-Based Outlier Detection, Proc. ACM SIGKDD, 2008.
- MARONNA, R., ZAMAR, R., Robust estimates of location and dispersion for high-dimensional data sets. Technometrics, 44(4), pp. 307–317, 2002.
- PAPADIMITRIOU, S., KITAGAWA, H., GIBBONS, P.B., FALOUTSOS, C., LOCI: Fast outlier detection using the local correlation integral. Tech. Rep. Intel Research Lab, IRP-TR-02-09, July 2002.
- PENA, D., PRIETO, F., Multivariate outlier detection and robust covariance matrix estimation. Technometrics, 43(3):286–310, 2001.
- RAMASWAMY, S., RASTOGI, R., KYUSEOK, S., Efficient algorithms for mining outliers from large data sets. Proc. ACM SIGMOD 2000, pp. 427–438, 2000.
- ROUSSEEUW, P.J., DRIESSEN, K.V., A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3), pp. 212–223, 1999.
- ROUSSEEUW, P.J., Van ZOMEREN, B.C., Unmasking Multivariate Outliers and Leverage Points, Journal of the American Statistical Association, 85, pp. 633-639, 1990.
- SHAWNE-TAYLOR J., CRISTIANI N., Kernel methods for pattern analysis. Cambridge, 2005.
- SHYU, M.-L., CHEN, S.-C., SARINNAPAKORN, K., CHANG, L., A novel anomaly detection scheme based on principal component classifier. Proc. ICDM 20003, pp. 353-365, 2003.
- SU, L., HAN, W., YANG, S., ZOU, P., JIA, Y., Continuous adaptive outlier detection on distributed data streams. In HPCC, LNCS 4782, pp. 74–85, 2007.
- SUBRAMANIAM, S., PALPANAS, T., PAPADOPOULOS, D., KALOGERAKI, V., GUNOPULOS, D., Online outlier detection in sensor data using non-parametric models. Proc. VLDB 2006, pp. 187–198, 2006.
- TANG, J., CHEN, Z., FU, A.W.-C., CHEUNG, D.W.-L., Enhancing Effectiveness of Outlier Detections for Low Density Patterns. Proc. PAKDD 2002. LNAI 2336, 2002.
- ZHANG, J., GAO, Q., WANG, H., Spot: A system for detecting projected outliers from high-dimensional data streams. Proc. ICDE 2008, pp. 1628–1631, 2008.

References

Missing Values

- ACUNA, E., RODRIGUEZ, C., The treatment of missing values and its effect in the classifier accuracy. Classification, Clustering and Data Mining Applications, Springer-Verlag, pp. 639-648, 2004.
- BATISTA G., MONARD, M.C., An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence 17, pp. 519-533, 2003.
- DEMPSTER, Arthur P., LAIRD, Nan M., RUBIN, Donald B., Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 39, 1–38, 1977.
- FAN, Wenfei, GEERTS, Floris. Relative Information Completeness, PODS'09, 2009.
- FARHANGFAR, A., KURGAN, L., DY, J., Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, 41, 3692-3705, 2008.
- FENG, H.A.B., Chen, G.C., Yin, C.D., Yang, B.B., Chen, Y.E., A SVM regression based approach to filling in missing values. Knowledge-Based Intelligent Information and Engineering Systems (KES05). LNCS 3683, pp. 581-587, 2005.
- HUA, Ming, PEI, Jian. Cleaning Disguised Missing Data A Heuristic Approach, Proc. KDD 2007.
- LI, D., DEOGUN, J., SPAULDING, W. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method. Rough Sets and Current Trends in Computing. LNCS 3066, 2004.
- LITTLE, R. J. A., RUBIN, D. B., Statistical Analysis with Missing Data. New York John Wiley Sons, 1987.
- Mc KNIGHT, P. E., FIGUEREDO, A. J., SIDANI, S., Missing Data A Gentle Introduction. Guilford Press, 2007.
- PEARSON, RONALD K., The problem of disguised missing data. SIGKDD Explorations 8(1) 83-92, 2006.
- SCHAFER, J. L., Analysis of Incomplete Multivariate Data, New York Chapman and Hall, 1997.
- TIMM, H., DORING, C., KRUSE, R., Different approaches to fuzzy clustering of incomplete datasets. International Journal of Approximate Reasoning, 35, 2003.
- WU, C.-H., WUN, C.-H., CHOU, H.-J., Using association rules for completing missing data. Proc. Hybrid Intelligent Systems (HIS’04), pp. 236–241, 2004.

References

Missing Values

- Allison, Paul D. (2002). Missing Data: Series: Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage Publications.
- Yuan, Y.C., 2000. Multiple imputation for missing data: concepts and new development. In: Proceedings of the Twenty-fifth Annual SAS Users Group International Conference. SAS Institute, Paper No. 267.
- Allison, P. D. 2000. Multiple Imputation for Missing Data: A Cautionary Tale. In Sociological Methods & Research, Vol. 28, No. 3, 301-309 (2000)